Animal Population Estimation Using Mark-Recapture and Plant-Capture

A thesis submitted by
Richard Gormley
to the University of St Andrews
in application for the degree of
Doctor of Philosophy

November 26, 2010

1. Candidate's declarations:

I, ...RICHARD GORMLEY ..., hereby certify that this thesis, which is approximately ... .. words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in SEPTEMBER 2005 [month, year] and as a candidate for the degree of .PhD.. . in SEPTEMBER 2006 [month, year]; the higher study for which this is a record was carried out in the University of St Andrews between 2005 [year] and 2011 [year].

(If you received assistance in writing from anyone other than your supervisor/s):
~~I, ..., received assistance in the writing of this thesis in respect of [language, grammar, spelling or syntax], which was provided by ......~~

Date ..12/6/12 ... signature of candidate ....

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of ..Ph.D... in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date ...9/6/12 ... signature of supervisor ....

3. Permission for electronic publication: *(to be signed by both candidate and supervisor)*

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

*Add **one** of the following options:*

~~(i) Access to printed copy and electronic publication of thesis through the University of St Andrews~~.

or

~~(ii) Access to [all or part] of printed copy but embargo of [all or part] of electronic publication of thesis for a period of ... years (*maximum five*) on the following ground(s):~~

~~(*delete those not applicable*)~~
~~publication would be commercially damaging to the researcher, or to the supervisor, or the University;~~
~~publication would preclude future publication;~~
~~publication would be in breach of law or ethics~~

or

✓ (iii) Embargo on both [all ~~or part~~] of printed copy and electronic copy for the same fixed period of ..2 years (*maximum five*) on the following ground(s):

(*delete those not applicable*)
~~publication would be commercially damaging to the researcher, or to the supervisor, or the University;~~
publication would preclude future publication;
~~publication would be in breach of law or ethics~~

or

~~(iv) Permanent embargo of [all or part] of print and electronic copies of thesis (*permission will be granted only in highly exceptional circumstances*)~~.

Date ..12/6/12... signature of candidate ...... signature of supervisor ....

A supporting statement for a request for an embargo must be included with the submission of the draft copy of the thesis. Where part of a thesis is to be embargoed, please specify the part and the reasons.

For Katie and Miles

Abstract

Mark-recapture is a method of population estimation that involves capturing a number of animals from a population of unknown size on several occasions, and marking those animals that are caught each time. By observing the number of marked animals that are subsequently seen, estimates of the total population size can be made. There are various subclasses of the mark-recapture method called the Otis-class of models (Otis, Burnham, White & Anderson 1978). These relate to the assumed behaviour of the individuals in the target population.

More recent work has generalised the theory of mark-recapture to the so-called plant-capture, where a known number of animals are pre-inserted into the target population. Sampling is then carried out as normal, but with additional information coming from knowledge of the number of planted individuals.

The theory underpinning plant-capture is less well-developed than mark-recapture, with the difference on population estimation of the former over the latter not often tested. This thesis shows that, under fixed and random sample-size models, the inclusion of plants can improve the mean point population estimation of various estimators. The estimator of Pathak (1964) is generalised to allow for the inclusion of plants into the target population. The results show that mean estimates from most estimators, under most models, can be improved with the inclusion of plants, and the sample standard deviations of the simulations can be reduced. This improvement in mean point population estimation is particularly pronounced when the number of animals captured is low.

Sample coverage, which is the proportion of distinct animals caught during sampling, is also often sought by practitioners. Given here is a generalisation of the inverse population estimator of Pathak (1964) to plant-capture and a proposed new inverse population estimator, which can be used as estimates of the coverage of a sample.

Acknowledgements

I would like to acknowledge my supervisors, Dr Ian Goudie and Professor Peter Jupp, for their help and support throughout this work. Ian, thank you for introducing me to mark-recapture and for endlessly being there for me when I sought assistance. I have developed so much under your supervision and I am truly grateful. Peter, I really appreciate the input that you gave me and for showing such genuine interest in my work. I am still trying to emulate your ability to go about your work with a continual smile on your face.

I would like to extend my deep gratitude to the Engineering and Physical Sciences Research Council (EPSRC) for funding this work, which made it possible for me to have such an enjoyable and eventful chapter of my life.

I would like to thank the University of St Andrews and its many staff and students for being such a large and enjoyable part of my life. Having been at St Andrews for over 8 years, there are too many people to mention, but special thanks go to Jun-Jie Miao, Daniel Mintz and Caleb O'Loan for making the Maths basement a less isolated place.

I would like to thank my family, especially David, who blazed the trail for me. A big thank you also goes to Ma and Da for encouraging me along the way.

A special thanks goes to Bernie Zech for pushing me over the finishing line.

Finally, I would like to thank my wonderful wife Carolyn and our beautiful daughter Katherine Isla. You have both given me so much happiness throughout this time, and completing it would not have been possible without your continued support and love (and occasional much-needed distractions).

A special mention also goes to Miles Anderson. Although you came after the completion of this work, you have already in your first few months of life given me so much joy and motivation for the future.

Contents

Chapter 1

INTRODUCTION

1.1 Overview

This thesis investigates new methods of mark-recapture and plant-capture for estimating animal abundance in a closed population. For the population to be closed, no migrations, births or deaths can occur during sampling.

Using plant-capture can have theoretical advantages over mark-recapture, but these advantages may be offset by practical difficulties, so one purpose of this work is to determine whether including plants is beneficial. Plant-capture has been carried out annually in New York City since 2005 to provide an estimate of the number of homeless people resident there. The Homeless Outreach Population Estimate (HOPE) census (Hopper, Shinn, Laska, Meisner & Wanderling (2008)) is carried out on one night, with some volunteers acting as street dwellers throughout the evening, allowing an estimate to be made from a single sampling occasion.[1]

Mark-recapture is the process of capturing a number of animals from a population of unknown size N, which is to be estimated. The number caught may or may not be predetermined, depending on which sampling scheme is used. These captured animals are distinctively, and permanently, marked and released back into their population. After a reintegration period the capturing process is repeated, with the previously marked animals recorded and any unmarked animals uniquely marked. This process is repeated until the experimenter deems appropriate, or some predetermined stopping criterion is satisfied. The total number of sampling occasions is denoted by t. At the end of the process various statistics (see §1.4) are recorded, some or all of which can be used for an estimate of abundance. Further information about the assumptions required for valid estimation can be found in Otis et al. (1978, pp 9-10) or Begon (1979, pp 8-9), and both provide tests of these assumptions. A test for closure can be found in the CloseTest program, given by Stanley

[1] Reports from previous surveys can be found at
http://www.nyc.gov/html/dhs/html/statistics/statistics.shtml

& Richard (2005).

Plant-capture uses a similar approach, but differs in that, before the first capturing occasion has taken place, a known number, R, of marked animals of the same class is added to the resident population. Sampling from the augmented population then proceeds as above. Plant-capture is a natural extension of mark-recapture, in that plant-recapture can be thought of as mark-capture with the first sampling occasion already having taken place. The j^{th} sampling occasion of mark-recapture $(j = 2, \ldots, t)$ thus coincides with the $(j - 1)^{th}$ sampling occasion in a plant-capture scenario with $t - 1$ samples. A subtle difference, however, is that, in mark-recapture situations, the number of plants, R, becomes random rather than fixed, which it is under plant-capture trials.

For both methods, as the focus here is only on closed populations, sampling usually takes place over a short period of time. This is consistent with Williams, Nichols & Conroy (2002, p. 331), which states:

> "A short period for the investigation increases the likelihood that the population remains closed to gains and losses over the period of sampling."

The rest of this chapter is dedicated to giving first a brief history of mark-recapture and plant-capture, defining the main variables and parameters that are required throughout this work and giving a brief description of the various models used throughout the next few chapters.

1.2 Background

The idea of marking animals can be traced back to the middle of the 17th century when Sir Francis Bacon tied ribbon on salmon and saw which of them returned upstream later, (Cormack (1968)). The use of mark-recapture to estimate population size, however, dates from the turn of the 20th century.

The earliest estimator was based on taking two samples and equating the proportion of animals caught in the first sample to the proportion of marked animals in the second sample. This is commonly referred to as the Petersen estimator, or as the Lincoln-Index by biologists, about which more is given later. The use of the ratio method can be traced back to the late 17th century when Graunt estimated the population of London, and the late 18th century, when Laplace estimated the population of France, (Cormack (1968)).

Mark-recapture estimation increased in popularity in 1938 with the paper by Zoe Schnabel (1938). She allowed for multiple recapturing occasions, and for the recapture probability to vary between capture occasions. This is referred to as the time-heterogeneous or time-dependent model, or simply the Schnabel census.

Since this time, mark-recapture theory and applications increased significantly[2] and continues to expand. For a more detailed history of mark-recapture, see Buckland, Goudie & Borchers (2000). For an overview of current areas of research on mark-recapture models, see Amstrup, McDonald & Manly (2005) or Pollock (2000).

Plant-capture is a newer area in which there has been interest recently, motivated originally by a desire by practitioners to generate population estimates from a single capturing occasion (Laska, Meisner & Siegel (1988, 1989)).

The methodology for such an estimator can be traced back to Rupp (1966), who gives an abundance estimator when there are "...two or more *kinds* of individuals in a population at Time 1 [before the first sample] ...". These kinds of individuals can be taken to be planted and target animals, since this group classification does not change throughout the trial. Rupp's (1966) abundance estimator requires an addition or removal of a known number of one or more types of animal at an intermediate point in the sampling procedure, whereas the work in this thesis considers populations of both the plant and target animals to be constant throughout the trial.

The United States Census Bureau used plant-capture methodology to estimate the number of homeless people in the 1990 U.S. census. For this, Laska & Meisner (1993) gives the likelihood function, and point and interval estimates for plant-capture. Papers by Goudie (1995), Yip (1996), Martin, Laska, Hopper, Meisner & Wanderling (1997), Goudie, Pollock & Ashbridge (1998), Goudie & Ashbridge (2000), Goudie, Jupp & Ashbridge (2007) and Ashbridge & Goudie (2009) have developed the subject further.

Mark-recapture was first put into a Bayesian framework by Freeman (1972, 1973), where he estimated population size under a sequential recapture framework. Castledine (1981) sought point and interval estimates of N under certain models using Beta priors for the capture probabilities. Smith (1991) used Bayes, empirical Bayes and Bayes empirical Bayes methods to compute point and interval estimates of N in the Schnabel census. George & Robert (1992) used Gibbs sampling to provide point population estimates. Reversible jump Markov chain Monte Carlo methods were used by King & Brooks (2001, 2002, 2008) to produce model averaged estimates. For an overview of the early expansion of Bayesian mark-recapture papers, one is referred to Schwarz & Seber (1999). Several Bayesian mark-recapture books have been published recently, firmly establishing Bayesian mark-recapture methods in ecology. Such books include McCarthy (2007), King, Morgan, Gimenez & Brooks (2009) and Link & Barker (2009).

[2] http://ncse.st-andrews.ac.uk/documents/posters/CapDataHistory.pdf

1.3 Closed mark-recapture

The work that I am carrying out will focus on closed animal populations under discrete-time sampling. In discrete-time sampling, sampling is carried out in distinct units, with a reintegration period between each sampling occasion. Thus, an animal can only be captured once at most in any sampling occasion. The alternative to discrete-time sampling is continuous-time sampling. With continuous-time sampling, it is often assumed that the capturing is carried out in one continuous interval, with the animals being immediately released upon tagging. In plant-capture situations, the population can also be assumed to be closed after the pre-marked animals are planted into the target population and allowed to cohabit. Discrete or continuous time recapturing can similarly be carried out.

Some useful notation is now defined, which is necessary for further discussion. When working under discrete time models, the word **sample** is taken to mean one process of capturing, and when **trial** is used, this is taken to mean the collection of all samples. So, one trial consists of t samples, where t is a fixed number chosen by the experimenter. A variety of models exist in mark-recapture and plant-capture, depending on how the capture probabilities relate between animals and between samples. For mark-recapture, the Otis-class of models, (Otis et al. 1978), which is an extension of a set of models attributed to Pollock (unpublished dissertation), attempts to account for various deviations from the equal-catchability model, denoted as model M_0, that are plausible in practise. The equal-catchability, or homogeneous, model assumes that there is an equal probability of capturing any animal in any sample. Variations from this include differing probabilities between animals in a particular sample but staying constant for each animal over samples, known as model M_h, differing probabilities between captured and uncaptured animals throughout the trial, referred to as model M_b, and model M_t, which assumes that in any sample every animal has the same probability of capture, but this probability differs between samples. The remainder of the Otis-class of models are the combinations of the three models, namely $M_{bh},\ M_{tb},\ M_{th}$ and M_{tbh}.

A model not considered by Otis *et al.* is the case when a predetermined number of animals are caught on each sample, which shall be referred to here as the fixed sample-size model, model M_f.

This work focuses mainly on plant-capture under model M_{tp}, where the p denotes working under the knowledge that planted animals are present. In cases where more than one subscript is used, the plant subscript will be placed in the final position. Thus, the model associated with time-dependent probability including plants will be referred to as the M_{tp} model, and so on.

As mark-recapture literature has expanded so much in the past 50 years, it is not possible to cover all areas in this thesis. Some notable areas that are not included are open population mark-recapture, models with individual heterogeneity and behavioural variations in capture probabilities and model selection.

For open population modelling, one is referred to Chapters 2 and 5 of Amstrup et al. (2005) and the many references therein.

For individual heterogeneity, Burnham & Overton (1978) proposed the jackknife method, which is the method employed by the program CAPTURE. The difficulties of estimating population size when animal heterogeneity is present are discussed in great detail in Link (2003).

For populations assumed to have a behavioural response to capture, Cormack (1989) proposed log-linear models and Lloyd (1994) proposed the martingale method.

Pledger (2000) gave a unified maximum likelihood framework to enable the fitting of all eight models given by Otis et al. (1978). She also provided a model selection procedure for choosing between the models. Other model selection methods have been proposed by Burnham, White & Anderson (1995) and Buckland, Burnham & Augustin (1997).

1.4 Statistics

The notation used in the mark-recapture literature is not standardised. Cormack (1968, p. 457) gives a table of the different notation used up until that point by various authors, which serves as a useful cross-reference. However, the primary reference for mark-recapture notation that one should use for this thesis is the brief list that is given below. Note that when subscripts are used, the letter i ($i = 1, \ldots, N$) will be used to indicate animal number and j ($j = 1, \ldots, t$) will be used to represent sample number.

N = The total (usually unknown) number of target animals in the target population.

R = The number of animals planted into the target population, $R \geq 0$.

t = The pre-chosen number of samples.

p_{ij} = The probability of capturing the i^{th} animal in the j^{th} sample.

z = The total observed number of captures observed (including plants when they are present) during the trial.

x = The number of distinct animals observed from the target population captured in the trial.

$\mathbf{n}$ = $(n_1, \ldots, n_t)$, where n_j is the number of animals (including plants when they are present) caught in the j^{th} sample. (Hence $0 \leq n_j \leq N + R$).

$\mathbf{m}$ = $(m_1, \ldots, m_t)$, where m_j is the number of marked animals caught in the j^{th} sample.

f_k = The number of animals caught exactly k times, $k = 0, \ldots, t$.

Note that some of the statistics and parameters differ when plants are introduced. As before, the parameter N and statistic x of X refer only to the target population, excluding plant captures. However, the other parameters and statistics given above, namely p_{ij}, z, $\mathbf{n}$, $\mathbf{m}$ and f_k now include the target and planted populations. Other, less-frequently used parameters and statistics will appear as they are required.
Another concept that is used widely is that of **coverage**. Sample coverage, C, is defined as being the ratio of the sum of the capture probabilities of the animals captured in the trial to the sum of the capture probabilities of all the animals (Chao, Lee & Jeng (1992)). This is given by

$$C = \frac{\displaystyle\sum_{i=1}^{N} p_i I[\text{the } i^{th} \text{ animal is captured}]}{\displaystyle\sum_{i=1}^{N} p_i}, \tag{1.1}$$

where I is the standard indicator function, equalling 1 if the i^{th} animal is caught, and 0 otherwise. When there is no heterogeneity between animals, however, the capture probability in any sample is the same for all animals. Thus, in homogeneous cases, (1.1) can be simplified to

$$C = \frac{x}{N}. \tag{1.2}$$

In practice the true population, N, and the capture probabilities, p_i are not known, so coverage must be estimated. Some such estimators are given in Chapter 6.

1.5 Scenarios

Discrete-time mark-recapture sampling generates observations that can be presented in an $(N + R) \times t$ matrix D, containing only 1s and 0s, where $d_{ij} = 1$ if the i^{th} animal is caught in the j^{th} sample, and $d_{ij} = 0$ otherwise, as first defined by Hammersley (1953). In the case of plant-capture a typical matrix will look like the following:

$$D = \begin{array}{l} \qquad\quad t \\ \left.\begin{matrix} 0 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{matrix}\right\} x \text{ rows} \\ \left.\begin{matrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{matrix}\right\} N - x \text{ rows} \\ \left.\begin{matrix} 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{matrix}\right\} R \text{ rows.} \end{array} \tag{1.3}$$

Thus, without loss of generality, the matrix can be ordered in such a way that the first x rows represent animals from the target population that have been caught at least once, and so each must contain at least one 1. The next $N - x$ rows contain only 0's, but this dimension is generally not known in practice. The final R rows are animals that are known to be present, but are not necessarily seen. Whether there is any relationship between the rows or columns depends on which model is applicable. The models that I have focussed on are detailed below.

Model	No. of parameters	Description
M_{fp}	1	Capture probability is not a parameter and so the only unknown parameter is N.
M_p	2	A constant capture probability, p, and population size, N, are assumed to be the only parameters.
M_{tp}	$t+1$	Capture probability assumed to differ in each sample, p_j, and population size, N is unknown, making $t+1$ assumed parameters.

Table 1.1: A brief summary of the number of assumed parameters for each model considered in this thesis.

1.6 Models

The details in the following subsections are written with the assumption of working under plant-capture scenarios, which subsume the corresponding mark-recapture models. If R is set to zero, one can get mark-recapture results explicitly. When animals are planted into a closed population, it is a requirement in the models given below that they behave in exactly the same way as the resident population. Thus, in this thesis, it is assumed that they integrate without rejection, and become, in homogeneous cases, as equally catchable as the target population.

If behavioural or animal heterogeneity is allowed for, this assumption is not required. Under behavioural models, planted animals can be assumed to have a different capture probability from the first sample onwards to that of the target population animals. Behavioural or animal heterogeneity are beyond the scope of this thesis, however.

1.6.1 Model M_{fp}

Model M_{fp} is not included in the Otis-class of models, but is one of the older models in mark-recapture history, and is covered in detail by Seber (1982). With this model, the number of animals captured in each sample is fixed before sampling begins. Thus, regardless of the catch-effort required, the specified number of animals should be caught in each sample, making the capture probability parameter redundant. This means that the only parameter is the unknown population size, N, as shown in Chapter 2.

1.6.2 Model M_p

Model M_p, the homogeneous case, is the simplest scenario when considering plant-capture models that are generalisations of the Otis-class of mark-recapture models. This model assumes that there are no behavioural differences between animals in any particular sample and so the probability of capturing any animal is the same, and it also assumes that this probability of capture remains constant between samples. It may seem unlikely in practice for such strict conditions to hold, but, nevertheless, this model yields some interesting results for many situations. This model provides a benchmark, giving a measure of the extent of the heterogeneity or behavioural effects present in the population when compared to the other models of the Otis-class.

This model has only two parameters, namely N and p, where p represents a constant probability of capture of any animal, $i = 1, \ldots, N + R$, in any sample $j = 1, \ldots, t$.

1.6.3 Model M_{tp}

Model M_{tp} introduces more reality into plant-capture sampling in that, whilst still assuming a common probability of capture for all animals in any particular sample, it assumes that the probability of capture differs between samples. Thus, this model can more realistically replicate real life scenarios, when assuming that the probability of capture remains constant over time can be hard to justify. One such situation is if the weather is different between subsequent samples, then the likelihood of capture on different days could be altered. Some examples are given by White, Anderson, Burnham & Otis (1982, p. 52) and Arnason, Kirby, Schwarz & Irvine (1996). Most of the work in this thesis has focussed on Model M_{tp}. This model has $t + 1$ parameters, namely N and p_j for $j = 1, \ldots, t$.

1.6.4 Model M_{ct}

Some reference will also be made to continuous-time models, where the notation adopted here will be to use a subscript, c, in the first position for a model corresponding to cases that are in continuous-time. Thus, when referring to a continuous-time, time-dependent model, the notation will be M_{ct}.

Continuous-time sampling differs from discrete-time sampling in that sampling is carried out continuously until a predetermined time or number of animals has been reached. Animals that are caught are tagged and released with the assumption that the time they were in captivity was negligible and had no significant effect on the capture probabilities in terms of time. Thus, if sampling is carried out in the interval $(0, \tau)$, then it is assumed that the sightings of any animal occur according to a Poisson process with rate $\lambda_i^*(t)$, which can be made a time-varying function defined in $(0, \tau)$, (see Hwang & Chao (2002)). The continuous time analysis can also be applied to behavioural models, by declaring

$$\lambda_i^*(t) = \begin{cases} \lambda_i \alpha(t) & \text{for previously uncaptured animals} \\ \phi \lambda_i \alpha(t) & \text{for previously captured animals.} \end{cases}$$

For M_{ct} scenarios, set $\lambda_i = 1$ and $\phi = 1$, removing inter-animal heterogeneity and behavioural effects. By using the parameters given in Amstrup et al. (2005) one can get all continuous time equivalents of the Otis et al. (1978) class of estimators. The only model mentioned in this work is the homogeneous model M_c. Under model M_c it is assumed that there is a constant rate parameter λ. By standard results, the total number of captures, Z, by time τ has a Poisson distribution with mean $N\lambda\tau$, given by

$$p(Z = z) = \frac{e^{N\lambda\tau}(N\lambda\tau)^z}{z!} \qquad\qquad z = x, x+1, \ldots. \tag{1.4}$$

Given $Z = z$, we have, by the classical occupancy distribution (c.f. Goudie et al. (1998)),

$$p(X = x|Z = z) = \frac{(N)_x}{N^z} S(x, z) \qquad x = 1, 2, \ldots, \min(N, z), \tag{1.5}$$

where $S(x, z) = \frac{1}{x!}\sum_{k=1}^{x}(-1)^k\binom{x}{k}(x-k)^z$, a Stirling Number of the Second Kind. Thus, the continuous-time joint probability distribution for X and Z for model M_c is

$$p(X = x, Z = z) = \frac{(\lambda\tau)^z\, e^{-N\lambda\tau}}{z!}\binom{N}{x}\sum_{k=0}^{x}(-1)^k\binom{x}{k}(x-k)^z \tag{1.6}$$
$$x = 0, 1, \ldots, N$$
$$z = x, x+1, \ldots.$$

This can then be used to find the maximum likelihood estimator for M_c, as (X, Z) is sufficient for (N, λ). As shown by Craig (1953), the M_c MLE is the solution of

$$\sum_{k=1}^{x}\frac{1}{N-k+1} = \frac{z}{N}. \tag{1.7}$$

Goudie et al. (1998) give a generalisation, for a fixed time τ, to the plant-capture equivalent model.

Another nice result, given by Lin & Chao (2005, p. 95), is that the MLEs for model M_{ctp} and model M_{tp} are asymptotically equivalent as $t \to \infty$.

1.7 Thesis summary

In Chapter 2, some work is carried out under models M_f and M_{fp}. For the mark-recapture model M_f, a comparative study is carried out between an estimator of Pathak (1964) and the commonly used estimators from the literature, comparing their means and standard deviations. For the plant-capture model M_{fp}, some theory is given to allow for the inclusion of plants. A generalised Pathak estimator is also derived under model M_{fp}, and compared against others from the literature. It is of interest to determine whether plants improve the estimate of population size sufficiently to compensate for the difficulties in satisfying the additional assumptions required.

In Chapter 3, the models M_0 and M_p are briefly covered. A lot of work has already been published for these models (see Ashbridge (1998), Goudie & Ashbridge (2005) and Ashbridge & Goudie (2009)), but the work contained here is intended to illustrate the performance of the generalised Pathak estimator under these models, which are models outside of its derivation

In Chapter 4, working under models M_t and M_{tp}, some work is also carried out to relate the generalised Pathak estimator to a class of estimators for restricted range Factorial Series Distributions, FSDs, as defined by Berg (1974). For the mark-recapture model M_f, Berg gives an equation for an unbiased estimator, conditional on the total number of captures, z, exceeding the total population size, N. He also offers an equation for the bias of the estimator when this condition is not satisfied. This could help to ease the computational difficulties that occur for the generalised Pathak estimator.

Chapter 5 is devoted to the specific case of mark-recapture and plant-capture where there are very few captures throughout the trial. This is known as sparse data mark-recapture. Under mark-recapture trials it has long been acknowledged that the reliability of an abundance estimate decreases if the recapture rate is low (see Borchers, Buckland & Zucchini (2002)).

Chapter 6 focusses on the estimation of sample coverage under plant-capture. It compares the estimators of Chao (1989) with an estimator proposed here.

Discussion of the previous chapters and some areas that should be considered in future work are given in the concluding chapter.

Chapter 2

MODEL M_{fp}

2.1 Introduction

Model M_f has been extensively discussed over the years. Darroch (1958, p. 345) showed that, for point and interval estimation of N, there is asymptotically no difference between fixed and random sample sizes, and states

> "...to estimate n $[N]$ as if the a_i $[n_i]$ are constants, when in fact they are not, is not a serious misrepresentation ...".

Under model M_f, Pathak (1964) gives the minimum variance unbiased estimator (MVUE) when $z \geq N$. One aim of this thesis is to use as few samples as deemed appropriate whilst obtaining a good estimate of N. Thus, as z increases proportionately with the number of samples, this condition will not always be met in this thesis. The effect of this condition not being met is addressed.

The generalisation of the unbiased estimator to give the MVUE under model M_{fp} is given in §2.3.2. I am unaware of any previous work on model M_{fp} explicitly, but Chapman (1952) allowed for the possibility of pre-marked individuals, without referring to them as plants.

What is offered in this chapter is the probability distribution of the sufficient statistic under model M_{fp} and the MLE for model M_{fp}. Also given is a generalisation of the Pathak estimator to allow for plants. To aid computation of this generalised Pathak estimator, equations (2.3) and (2.4) of Berg (1976) are generalised to model M_{fp}.

Finally, I present a comparative study of whether including plants improves estimation sufficiently to justify their inclusion, discuss what an optimal number of plants would be and whether the simpler to compute Pathak approximation estimator can be favoured over the Pathak estimator.

2.2 Probability Theory

A generalisation of a result given by Seber (1982, p. 132), to allow for the inclusion of plants, gives the distribution of the number of marked animals caught in each sample given $\mathbf{n}$. This makes use of the vector $\mathbf{m} = (m_1, \ldots, m_t)$, which represent the number of animals caught in the j^{th} sample $(j = 1, \ldots, t)$ that were previously marked, as given in §1.4. This differs from Seber (1982, p. 130) in that here, m_1 is not necessarily 0, since plants are inserted prior to the first sampling occasion. Also, we require $\mathbf{u} = (u_1, \ldots, u_t)$, where u_j represents the number of animals caught in the j^{th} sample $j = 1, \ldots, t$ that were previously unmarked. We also define the vector $\mathbf{M} = (M_1, \ldots, M_t)$ where $M_1 = R$ and $M_j = R + \sum_{k=1}^{j-1} u_k$, $\{j = 2, \ldots, t+1\}$, which represents the number of marked animals just before the j^{th} sample is taken. Thus, we get $M_{t+1} = x + R$. We also define N_A as being the total number of animals in the augmented population, *i.e.* $N_A = N + R$. Using these, we get a product of hypergeometric distributions (c.f. Bishop, Fienberg & Holland (1975, §13.5)) for $\mathbf{M}$, namely

$$\begin{aligned} f(m_1, \ldots, m_t | \{n_j\}) &= \prod_{j=1}^{t} \frac{\binom{M_j}{m_j}\binom{N_A - M_j}{u_j}}{\binom{N_A}{n_j}} \\ &= \frac{N!}{(N-x)!} \prod_{j=1}^{t} \frac{\binom{M_j}{m_j}}{\binom{N_A}{n_j}(u_j)!}, \end{aligned} \tag{2.1}$$

for $m_1 = 0, \ldots, \min(n_1, R)$ and $m_j = 0, \ldots, \min(n_j, R + \sum_{k=0}^{j-1} n_k)$, $j = 2, \ldots, t$. The Neyman-Fisher factorisation theorem then gives that X is sufficient for N_A, or, equivalently, N, the size of the target population.

Thus, we seek to establish the probability distribution of X, (c.f. Goudie & Gormley (in submission)). For this we will use the inclusion-exclusion principle (c.f. Johnson, Kotz & Kemp (2005, p. 432))

Inclusion-exclusion principle: Given N objects, suppose that $n(a)$ have property a, $n(b)$ have property b,..., $n(ab)$ have properties a and b,..., $n(abc)$ have properties a, b and c etc. Then the inclusion-exclusion principle states that the total number of objects with none of these properties is

$$\begin{aligned} N(\bar{a}\bar{b}\bar{c}\ldots) = \quad & N \quad -n(a) - n(b) - n(c) - \ldots \\ & + \quad n(ab) + n(ac) + \ldots \\ & - \quad n(abc) - \ldots \\ & + \quad \ldots \end{aligned}$$

■

Firstly we need to find the total number of ways to select all the t samples, such that n_j animals are caught in sample j $(j = 1, \ldots, t)$. This is the number of ways that the t samples can be chosen from the $N + R$ animals present. This is just

$$A(N, \mathbf{n}, R) = \prod_{j=1}^{t} \binom{N+R}{n_j}. \tag{2.2}$$

Also, the number of ways of getting x distinct captures from N animals is simply $\binom{N}{x}$. If we now, without loss of generality, rearrange the animals such that the x captured animals are ordered first, we get $K = A(x, \mathbf{n}, R) \subseteq A(N, \mathbf{n}, R)$ as the total number of ways that we can choose the samples from the $x + R$ observed animals.

We now need to find the total number of selections from K where all x animals are seen at least once. Let ω denote a subset of $\{1, 2, \ldots, x\}$ and let A_ω be such that when $i \in \omega$ the i^{th} of these x individuals is not seen in any of the t samples. Let $n(A_\omega)$ denote the number of selections from $A(x, \mathbf{n}, R)$ that have the property A_ω. The inclusion-exclusion principle gives the number of combinations in K in which all x animals are caught at least once as

$$\begin{aligned} A(x, \mathbf{n}, R) \quad & - \quad n(A_{\{1\}}) - n(A_{\{2\}}) - \ldots \\ & + \quad n(A_{\{1,2\}}) + n(A_{\{1,3\}}) + \ldots \\ & - \quad n(A_{\{1,2,3\}}) - \ldots \\ = \quad & \quad A(x, \mathbf{n}, R) - \binom{x}{1} n(A_{\{1\}}) + \binom{x}{2} n(A_{\{1,2\}}) - \ldots \\ = \quad & \quad x! a(x, \mathbf{n}, R), \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} a(x, \mathbf{n}, R) &= \frac{1}{x!} \Delta^x \left[A(N, \mathbf{n}, R) \right]_{N=0} \\ &= \frac{1}{x!} \sum_{k=0}^{x} (-1)^k \binom{x}{k} \prod_{j=1}^{t} \binom{R + x - k}{n_j}, \end{aligned} \qquad (2.4)$$

and where

$$\Delta^x \left[f(t) \right]_{\omega=0} = \sum_{k=0}^{x} \binom{x}{k} (-1)^k f(x + t - k) \qquad (2.5)$$

is the x^{th} forward finite difference ($x = \max_{j=1,\ldots,t}(n_j - R, 0), \ldots, \min(N, z)$), evaluated at $\omega = 0$. Effectively this generalises Pathak (1964) to include plants. It follows that the probability of X conditional on the n_js is given by:

$$p(x|\mathbf{n}) = \frac{(N)_x \, a(x, \mathbf{n}, R)}{A(N, \mathbf{n}, R)}, \qquad (2.6)$$

where $(N)_\nu = N(N-1)(N-2)\ldots(N-\nu+1)$ is a truncated factorial and X is defined between $x = \max_{j=1,\ldots,t}(n_j - R, 0), \ldots, \min(N, z)$.
This is of the form of a factorial series distribution, or FSD, with series function as defined by Berg (1974), and implies the existence of a unique unbiased estimator of N, which is given in §2.3. As this probability distribution is of closed form, with a single unknown parameter, it is possible to calculate the moments of an M_{fp} estimator exactly, without the need for simulation. Some such moments will be given in §2.5.

2.3 Estimators

2.3.1 Maximum Likelihood Estimator for M_{fp}

Using (2.6) we can determine the maximum likelihood estimate for model M_{fp}, since we have:

$$L(N; x) \propto \frac{(N)_x}{A(N, \mathbf{n}, R)}.$$

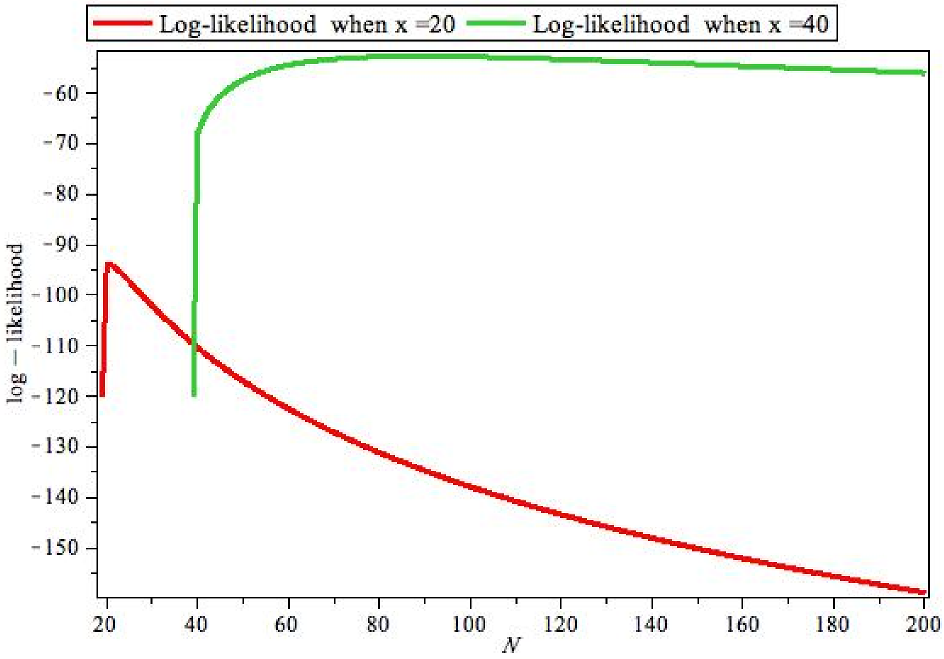


Figure 2.1: An example log-likelihood plot for a model M_f trial with $\mathbf{n}$={10,10,10,10,10}, $R = 0$ and $x = 20$ (given in red when the M_f MLE estimate is 20) and $x = 40$ (given in green, where the M_f MLE estimate is 89). Both plots have the arbitrary starting value of -120.

Taking logs of both sides gives the log-likelihood function under M_{fp} as:

$$\begin{aligned}\ell(N;x) &= \sum_{i=0}^{x-1}\log(N-i) - \log\left(\prod_{j=1}^{t}\binom{N+R}{n_j}\right) + c_1 \\ &= \sum_{i=0}^{x-1}\log(N-i) - \sum_{j=1}^{t}\log\left[(N+R)_{n_j}\right] + c_2 \\ &= \sum_{i=0}^{x-1}\log(N-i) - \sum_{j=1}^{t}\sum_{i=0}^{n_j-1}\log(N+R-i) + c_3 \qquad (2.7)\end{aligned}$$

where c_1, c_2 and c_3 are constants. From here, we use the fact that the likelihood is unimodal (c.f. Goudie & Gormley (in submission) or see Appendix A), and seek to find its maximum value. Figure 2.1 gives a plot of (2.7) for a trial with $\mathbf{n} = \{10, 10, 10, 10, 10\}$ and no plants present. Plotted are log-likelihood functions for two possible values of $x(= 20$ and $40)$, which show the shift in the mode as x increases. The log-likelihood remains unimodal, however.

No closed-form solution exists for this except for when $t = 2$, so a recursive method is used to determine the maximum likelihood estimate $\hat{N}_{\text{MLE}}$ of N. This is given as the largest $\aleph \in \{x, x+1, x+2, \ldots\}$ such that $L(\aleph; x) > L(\aleph - 1; x)$, or:

$$\begin{aligned}
\log(\aleph) - \log(\aleph - x) &> \sum_{j=1}^{t} [\log(\aleph + R) - \log(\aleph + R - n_j)] \\
\log\left(\frac{\aleph}{\aleph - x}\right) &> \log\left\{\frac{(\aleph + R)^t}{\prod_{j=1}^{t}(\aleph + R - n_j)}\right\} \\
\iff \qquad \aleph \prod_{j=1}^{t}(\aleph + R - n_j) &> (\aleph - x)(\aleph + R)^t. \qquad (2.8)
\end{aligned}$$

The maximum likelihood estimate is tested in §2.5 under various scenarios.
Profile likelihood confidence intervals can be calculated relatively easily for model M_{fp}, as the profile (log-)likelihood is equal to the (log-)likelihood in this model, as there is only one parameter. Thus, a confidence interval can be calculated without assuming normality, allowing non-symmetric intervals.
Let $\ell(\hat{N}_{\text{MLE}})$ be the value of the profile log-likelihood evaluated at $\hat{N}_{\text{MLE}}$ and $\ell(N)$ be the value of the profile log-likelihood evaluated at $N \in \{x, x+1, x+2, \ldots\}$. Thus, we have

$$W(N) = 2\left[\ell(\hat{N}_{\text{MLE}}) - \ell(N)\right] \dot{\sim} \chi^2_1. \qquad (2.9)$$

For a 95% confidence interval we use the fact that $\chi^2_{1;0.05} = 3.84$ to find the lower and upper tail values.

2.3.2 Generalised Pathak estimator

Pathak's (1964) estimator was designed to be unbiased for M_f but is here adapted to model M_{fp}. The results of Berg (1974) show that, under M_{fp} the estimator

$$\tilde{N}(x, \mathbf{n}, R) = x + \frac{a(x-1, \mathbf{n}, R)}{a(x, \mathbf{n}, R)} \qquad (2.10)$$

is an unbiased estimator of N when $z \geq N$. It is of interest to see how the generalised Pathak estimator performs when this condition fails. This is shown in §2.5.1.
The generalised Pathak estimator (2.10) can be shown to be the minimum variance unbiased estimator under model M_{fp} by noticing the equivalence of (2.10) to Berg's (1974) equation (2.14), and then using the results therein.
Pathak (1964, p. 79) states that his estimator is "difficult to compute" unless $n_1 = n_2 = \ldots = n_t = 1$ and $t \leq 50$. The difficulty lies in the computation of the a-coefficients, (2.4). Computation of the a-coefficients becomes increasingly difficult as the population size increases, as the terms of the coefficient can become too large for some computer programs to handle. Computation can be made simpler

by using a recursive formula for the a-coeffiicients, (2.4), which was given by Berg (1976, eq. (2.3)) for the $R = 0$ case and generalised here to the plant-capture case. We begin by defining $\mathbf{n}' = (n_1, \ldots, n_{t-1})$ and writing (2.6) as

$$p(x|\mathbf{n}) = \sum_{\nu=0}^{x} p(x-\nu|\mathbf{n}') \frac{\binom{R+x-\nu}{n_t-\nu}\binom{N-(x-\nu)}{\nu}}{\binom{N+R}{n_t}}, \tag{2.11}$$

for $x = \max_{j=1,\ldots,t}(n_j - R, 0), \ldots, \min(N, z)$. Setting (2.11) equal to (2.6) and rearranging for $a(x, \mathbf{n}, R)$ gives

$$a(x, \mathbf{n}, R) = \frac{1}{(N)_x} \prod_{j=1}^{t-1} \binom{N+R}{n_j} \sum_{\nu=0}^{x} p(x-\nu|\mathbf{n}') \binom{R+x-\nu}{n_t-\nu}\binom{N-(x-\nu)}{\nu}. \tag{2.12}$$

Observing that

$$p(x-\nu|\mathbf{n}') = \frac{(N)_{x-\nu} a(x-\nu, \mathbf{n}', R)}{A(N+R, \mathbf{n}', R)}$$

and simplifying (2.12) gives

$$a(x, \mathbf{n}, R) = \sum_{\nu=0}^{x} \frac{1}{\nu!} \binom{R+x-\nu}{n_t-\nu} a(x-\nu, \mathbf{n}', R). \tag{2.13}$$

Berg's (1976, eq. (2.4)) can be similarly generalised to the plant-capture case by firstly writing (2.6) as

$$p(x|\mathbf{n}) = \frac{R+x-(n_t-1)}{N+R-(n_t-1)} p(x|\mathbf{n}'') + \frac{N-x+1}{N+R-(n_t-1)} p(x-1|\mathbf{n}''), \tag{2.14}$$

for $x = \max_{j=1,\ldots,t}(n_j - R, 0), \ldots, \min(N, z)$, $\mathbf{n}'' = (n_1, \ldots, n_t - 1)$ and where $\mathbf{n}'' = \mathbf{n}'$ if $n_t = 1$. Setting (2.14) equal to (2.6), rearranging for $a(x, \mathbf{n}, R)$ and simplifying gives

$$a(x, \mathbf{n}, R) = \frac{R+x-(n_t-1)}{n_t} a(x, \mathbf{n}'', R) + \frac{1}{n_t} a(x-1, \mathbf{n}'', R). \tag{2.15}$$

To compensate for the difficulties in calculating his estimator, Pathak (1964, p. 79) also gives an approximation to his estimator, (2.10), using a ratio of what he calls differences of zeroes, given in Pathak (1961). The Pathak approximation, $\tilde{N}_{PA}$, is defined as

$$\tilde{N}_{PA} = \frac{C_x(z+1)}{C_x(z)}, \tag{2.16}$$

where the differences of zeroes are defined as

$$\begin{aligned} C_x(z) &= x^z - \binom{x}{1}(x-1)^z + \ldots (-1)^{x-1}\binom{x}{x-1} \\ &= x!S(z,x), \end{aligned} \tag{2.17}$$

where $S(z,x)$ is the Stirling number of the second kind with arguments z and x, given by

$$S(z,x) = \frac{1}{x!}\sum_{k=0}^{x}(-1)^k\binom{x}{k}(x-k)^z.$$

The improved computational power now available, however, means that the generalised Pathak estimator, (2.10), can be computed for non-unitary numbers of animals caught in each sample. The approximate estimator (2.16) is compared to the generalised Pathak estimator (2.10) below.

2.3.3 The M_p conditionally unbiased estimator

The generalised Pathak estimator was shown to be the MVUE under model M_{fp}, where there is only one sufficient statistic, X. In model M_p, there are two sufficient statistics, X and Z. This model is the focus of the next chapter. However, the Rao-Blackwellised Pathak estimator under model M_p gives the conditionally unbiased estimator (CUE) under model M_p. This M_p CUE is used here in a model outside of that for which it was derived. A detailed derivation will be given in the next chapter, but the estimator is given as

$$\tilde{N}_c = x + \frac{G(z,x-1,t,Rt)}{G(z,x,t,Rt)}, \tag{2.18}$$

where

$$\begin{aligned} G(z,x,t,Rt) &= \frac{1}{x!}\Delta^x\left[(Rt+\omega t)_z\right]_{\omega=0} \\ &= \frac{z!}{x!}\sum_{k=0}^{x}(-1)^k\binom{x}{k}\binom{Rt+xt-kt}{z}. \end{aligned} \tag{2.19}$$

The $G(z,x,t,Rt)$ coefficients are called the Gould-Hopper numbers (Gould & Hopper 1962).

This Rao-Blackwellised Pathak estimator was compared with (2.8) and (2.10), and the results are given in §2.5.

2.4 Computation

Under model M_{fp} it is possible to compute moments of the above estimators exactly, as the probabilities at all points in the support of the sufficient statistic, X, can be computed relatively simply. The first moment, the mean or expected value, is calculated by multiplying the particular estimator's value at $x \in X$ with $p(X = x)$, (2.6), and summing over all $x \in X$.

Results are given below of some computation carried out, where the assumption is that the same number of animals are caught in each sample. Estimates are given for trials with fixed population size, N $(= 10, 50 \text{ or } 100)$, both with and without the addition of a fixed number of plants, R $(= 0, 5, 10 \, \text{or} \, 15)$. For each choice of N and R, estimates are generated for almost every possible trial outcome, from 0 to N captures in each sample. These are plotted on graphs, with the true population and the line giving $p(z = x)$ for each trial overlaid.

Also given is some analysis to determine the optimal number of plants that the experimenter should use. Another comparison made is to evaluate the performance of the approximation to the Pathak estimator that is given in Pathak (1964) against the Pathak estimator.

In all of the computations outlined above, the restriction that there must be at least one recapture in the trial has been imposed on both the generalised Pathak estimator and the M_p CUE. This assumption was deemed appropriate to allow for a consistent comparison between them and the M_{fp} MLE, which requires this restriction in order to have a finite mean. Another restriction imposed on the two non-MLE estimators is to round their estimates to the nearest integer. This restriction is again to allow for a fairer comparison between all three estimators.

2.5 Results

The results section is broken down into several parts. Subsection 2.5.1 attempts to establish the optimal estimator under model M_{fp}. Subsection 2.5.2 then seeks to provide evidence for the optimal number of plants to improve the mean population point estimates for each estimator. Subsection 2.5.3 compares the Pathak estimator (2.10) (with $R = 0$) with its approximation (2.16), as given in Pathak (1964).

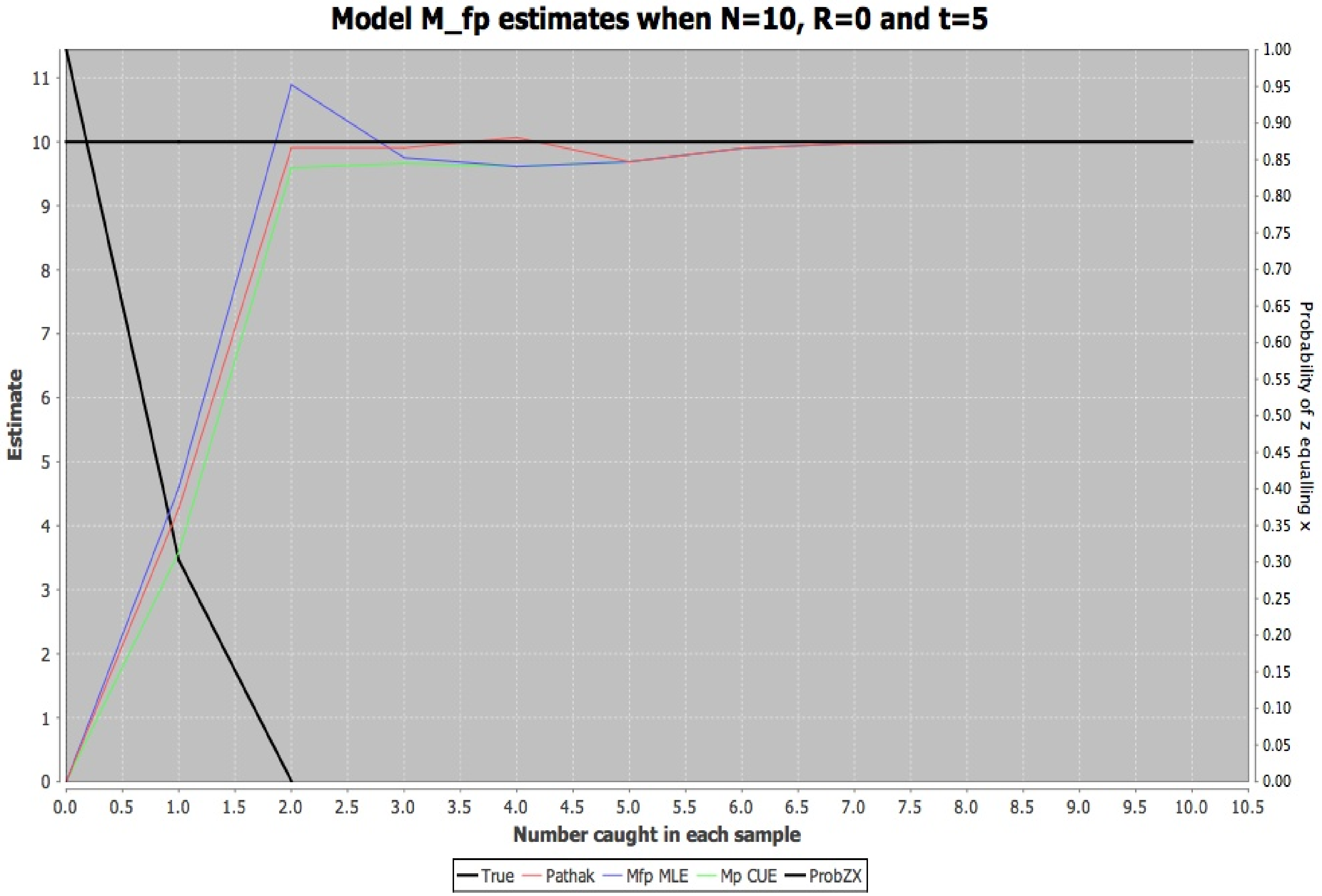


Figure 2.2: Plot of the generalised Pathak (red), M_{fp} MLE (blue) and M_p CUE estimates against the number of animals caught in each sample when $N = 10$ and $R = 0$, under model M_f. Overlaid is the true population size and the probability that $z = x$ for each trial (in black), which uses the right-hand axis.

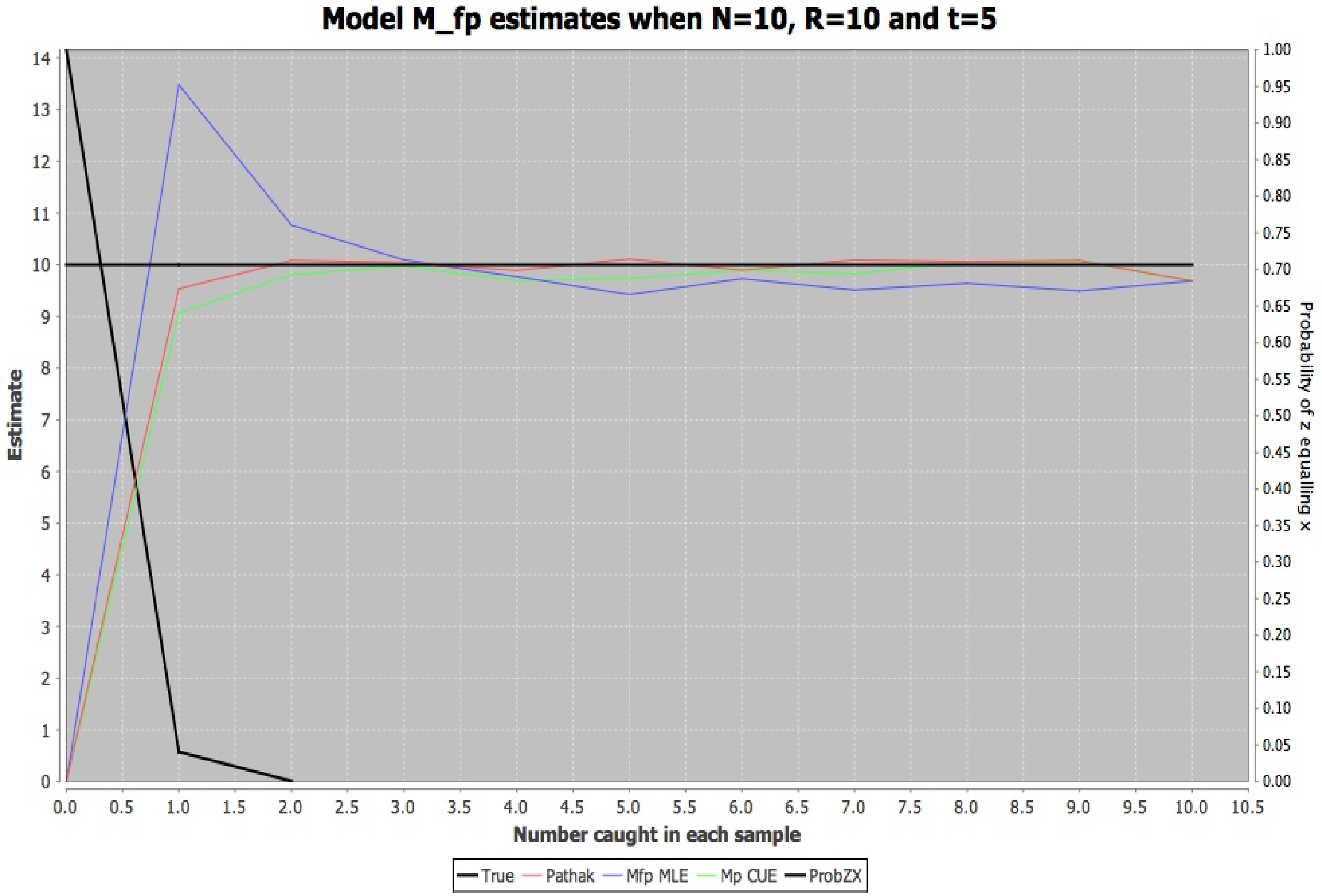


Figure 2.3: Plot of the generalised Pathak (red), M_{fp} MLE (blue) and M_p CUE estimates against the number of animals caught in each sample when $N = 10$ and $R = 10$, under model M_{fp}. Overlaid is the true population size and the probability that $z = x$ for each trial (in black), which uses the right-hand axis.

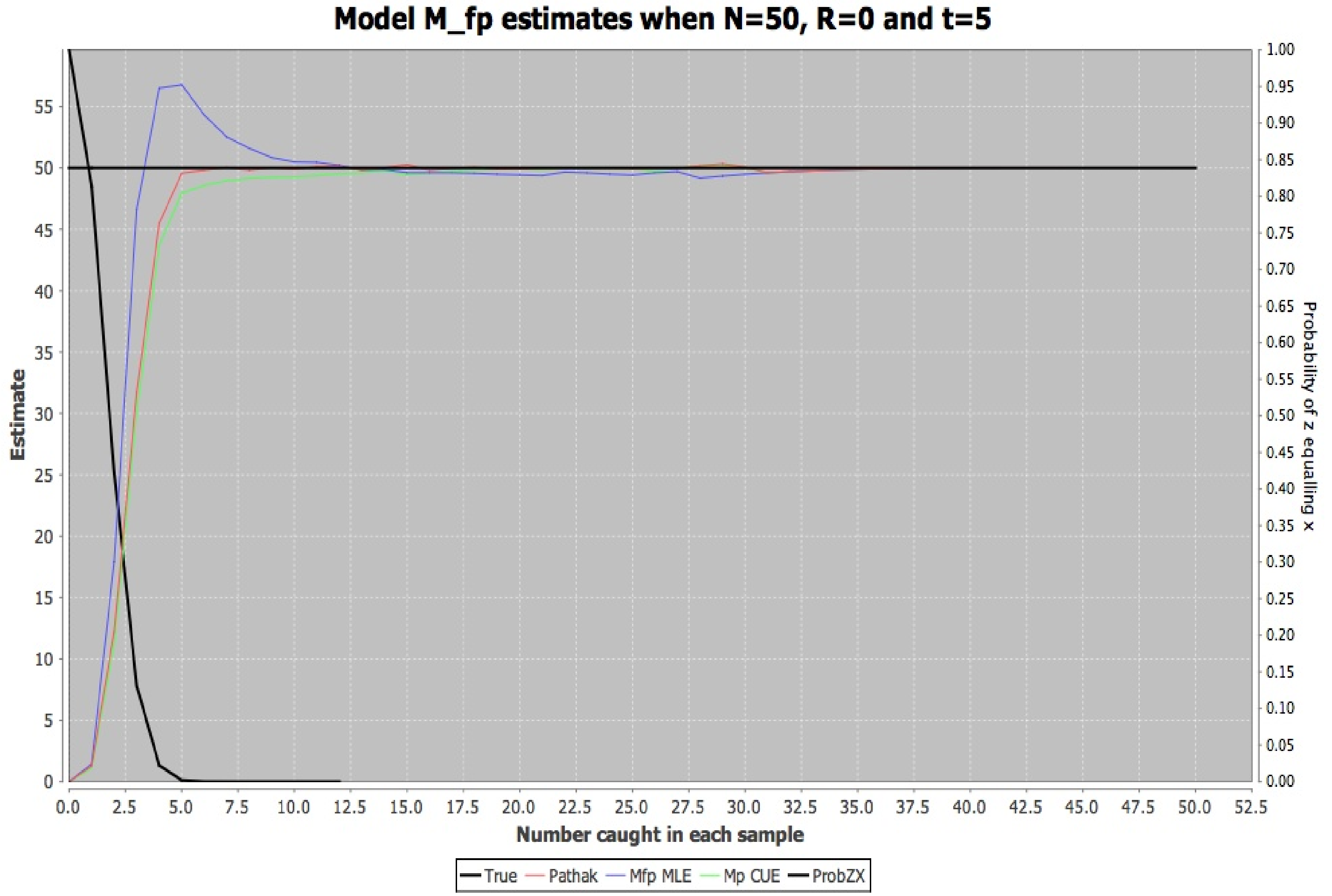


Figure 2.4: Plot of the generalised Pathak (red), M_{fp} MLE (blue) and M_p CUE estimates against the number of animals caught in each sample when $N = 50$ and $R = 0$, under model M_f. Overlaid is the true population size and the probability that $z = x$ for each trial (in black), which uses the right-hand axis.

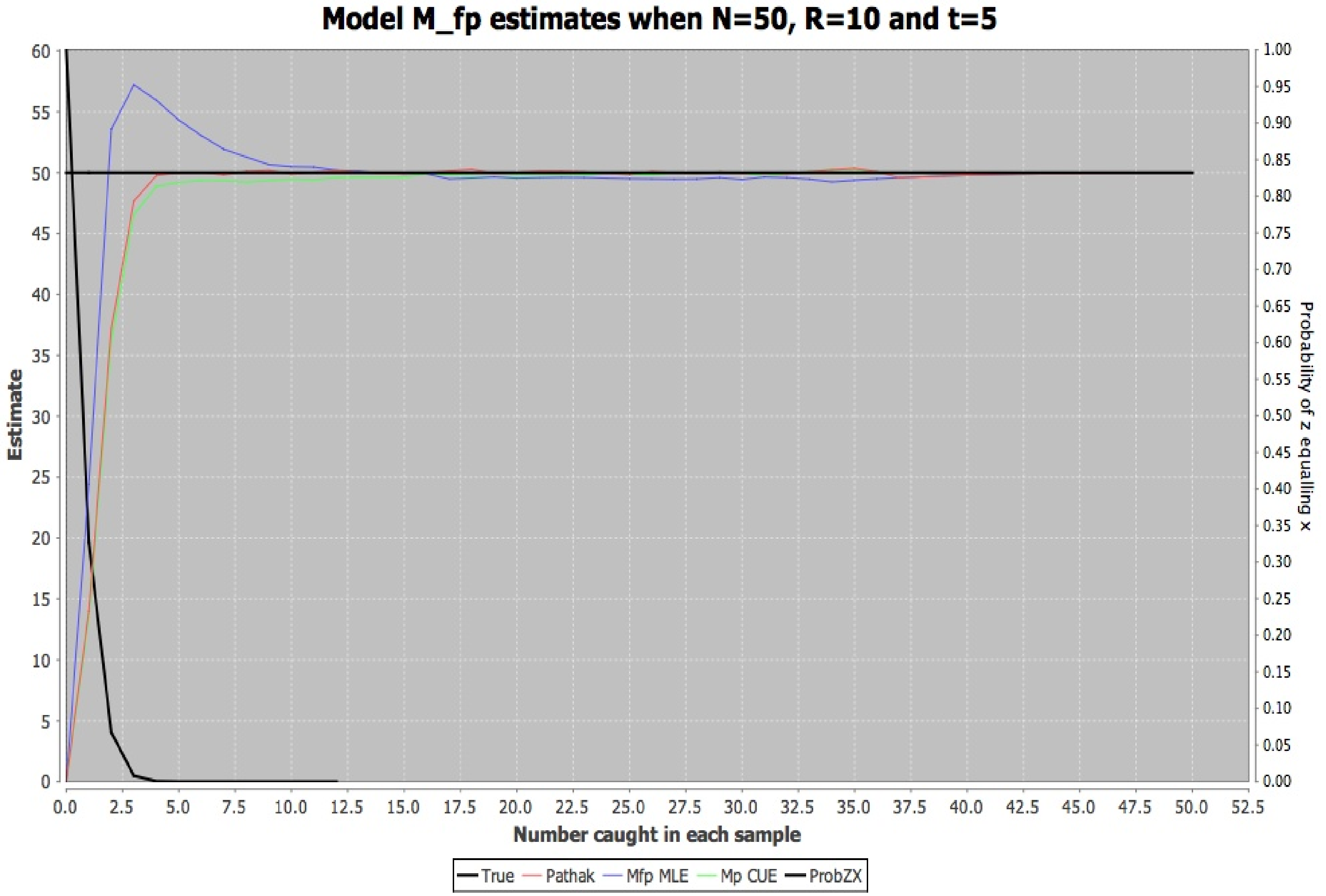


Figure 2.5: Plot of the generalised Pathak (red), M_{fp} MLE (blue) and M_p CUE estimates against the number of animals caught in each sample when $N = 50$ and $R = 10$, under model M_{fp}. Overlaid is the true population size and the probability that $z = x$ for each trial (in black), which uses the right-hand axis.

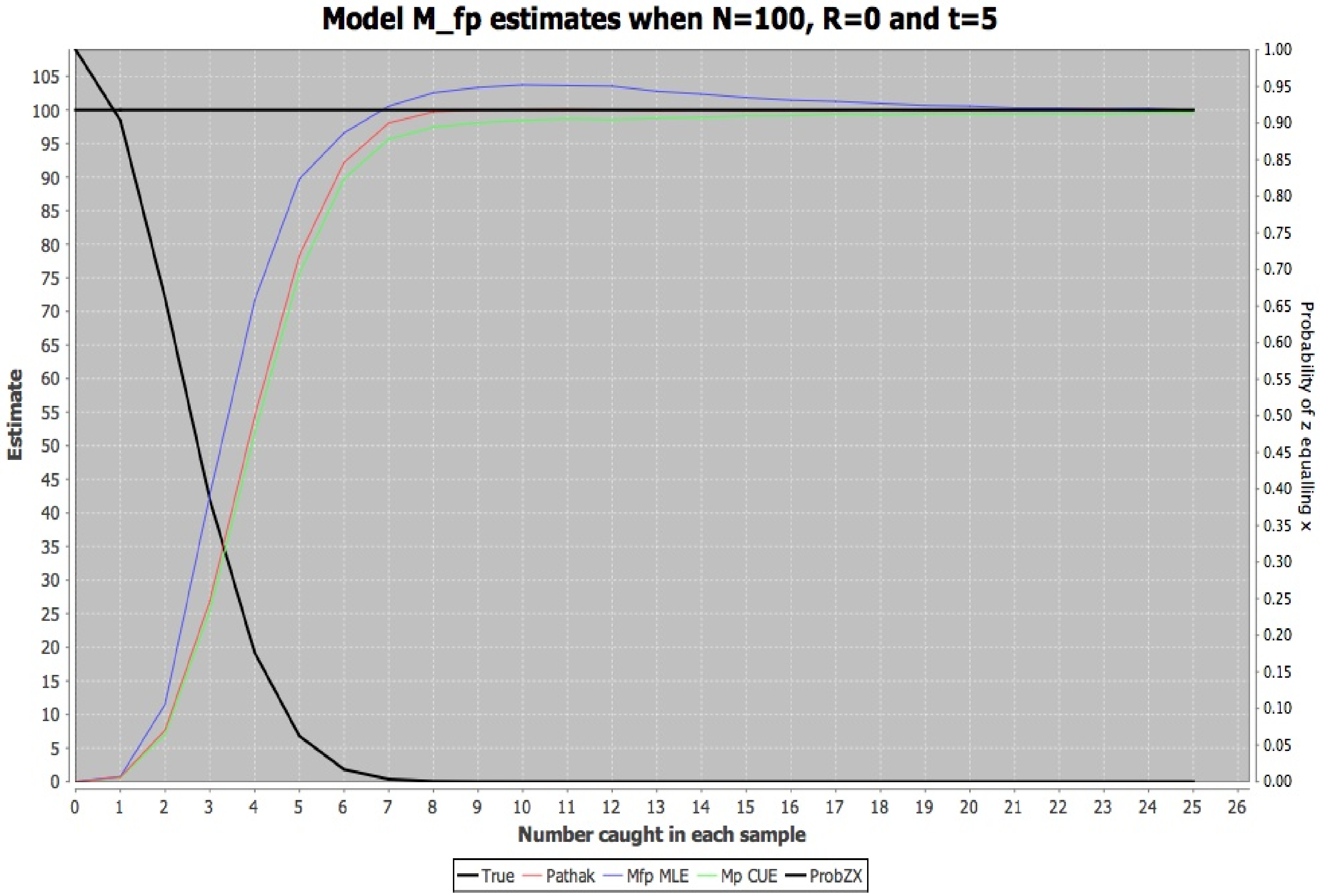


Figure 2.6: Plot of the generalised Pathak (red), M_{fp} MLE (blue) and M_p CUE estimates against the number of animals caught in each sample when $N = 100$ and $R = 0$, under model M_f. Overlaid is the true population size and the probability that $z = x$ for each trial (in black), which uses the right-hand axis.

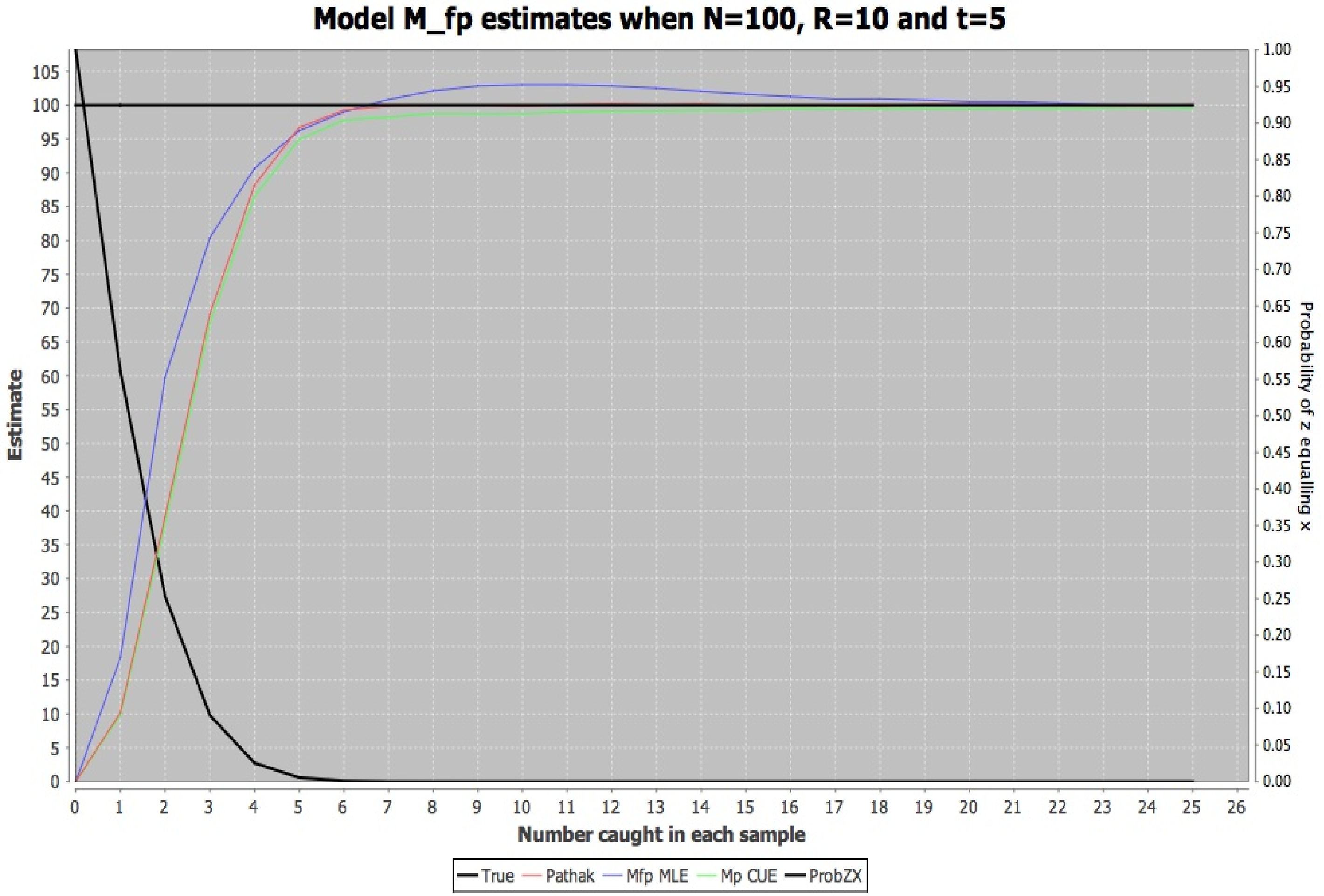


Figure 2.7: Plot of the generalised Pathak (red), M_{fp} MLE (blue) and M_p CUE estimates against the number of animals caught in each sample when $N = 100$ and $R = 10$, under model M_{fp}. Overlaid is the true population size and the probability that $z = x$ for each trial (in black), which uses the right-hand axis.

2.5.1 Optimal estimator

From Figures 2.2–2.7 it can be seen that the expected value of the M_{fp} MLE has the steepest gradient when starting at the origin, having an expected value larger than both of the other estimators when each n_j is small. However, it is also evident that this estimator, for at least 10% of the trials plotted, has an expected value that overestimates the true population, tending towards and then oscillating around the true population size. This is an undesirable feature of the M_{fp} MLE, as overestimation in practice can give falsely high predictions of abundance, which may understate the severity of population decline, for example.

It can be seen from Figures 2.2–2.7 that the expected values of the generalised Pathak estimator do not (within rounding) overestimate the true population size for any Z or $\mathbf{n}$ that has been modelled. It also provides almost unbiased estimates for most sample sizes modelled. Figures 2.2–2.7 also show that the condition $z \geq N$ can be relaxed, as the estimator is unbiased for some trials when this condition is not satisfied. Although the initial gradient is not quite as steep as the M_{fp} MLE for very small sample sizes, this difference is small. It can also be seen that the gradient for the generalised Pathak estimator is slightly greater than that of the M_p CUE. The mean of the M_p CUE has the shallowest gradient of the three estimators and it lags behind the generalised Pathak estimator in tending towards the true population size in each trial. It is evident, however, that mean estimates tend towards the true population size for many of the constant sample size trial scenarios as the sample sizes increase.

In Figures 2.6 and 2.7 the difference between the generalised Pathak estimator and the M_p CUE is particularly pronounced, with the M_p CUE approaching the true population size, $N = 100$, but struggling to attain it exactly. The generalised Pathak estimator, however, reaches 100 at $\{n_j = 8,\ j = 1, \ldots, t\}$ when $R = 0$ and with slightly fewer captures in each sample when $R = 10$. Also, once the generalised Pathak estimator mean estimate reaches N, then it rarely deviates from producing an unbiased expected value.

For all three population sizes simulated, it is evident that the gradient is steeper for all three estimators when 10 plants are included than when no plants are present. For the generalised Pathak estimator and the M_p CUE this always leads to improved mean estimates, but for the M_{fp} MLE it can lead to a trial producing an overestimated mean rather than an underestimated mean. Thus, for the M_{fp} MLE care must be taken to ensure that a sufficient number of animals are caught in each sample, although it is not straightforward to establish in practice what this number should be. This problem can be avoided by using either the generalised Pathak estimator or the M_p CUE.

N=50							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(2,2,2,2,2)	12.2512	11.0217	17.9132	16.9395	11.2850	10.1090
10	(2,2,3,2,2)	40.4810	21.1034	54.2548	31.9683	39.0576	20.5913

N=50							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(5,5,5,5,5)	49.5716	19.5460	56.7757	25.1724	47.9544	18.9295
10	(6,6,6,6,6)	50.0304	15.3832	53.0356	17.4893	49.3536	15.1237

N=50							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(10,10,10,10,10)	49.8815	7.8766	50.4909	8.4990	49.2826	7.8776
5	(11,11,11,11,11)	49.9988	7.5416	50.2682	7.7002	49.3870	7.2453
10	(12,12,12,12,12)	50.1829	7.1349	50.1862	7.2172	49.5874	6.9276

N=50							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(25,20,15,10,5)	49.9656	3.9198	49.9497	3.9562	51.2572	4.2806
5	(30,22,15,11,4)	50.3090	3.8586	49.4962	3.8165	52.3108	3.9296
10	(35,25,15,10,5)	50.1763	3.7281	49.5065	3.4624	52.3231	3.6908

N=50							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(25,25,25,25,25)	49.9570	1.5709	49.4269	1.1689	49.9570	1.5709
10	(30,30,30,30,30)	49.9595	1.5869	49.4255	1.1902	49.9595	1.5869

Table 2.1: Expected values from 1000 realisations of the estimators under various M_{fp} scenarios for a population, N=50.

N=100							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(5,5,5,5,5)	78.3102	36.3619	89.6988	36.2476	75.6684	34.9340
10	(6,5,6,5,6)	98.6520	44.5782	98.9818	26.8597	97.1603	43.6165

N=100							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(10,10,10,10,10)	100.0925	28.6556	103.7340	25.3362	98.4572	28.1246
10	(11,11,11,11,11)	100.0481	23.7294	102.9997	23.2279	98.9709	23.4471

N=100							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(20,20,20,20,20)	99.9725	11.2315	100.5807	11.6493	99.4046	10.9806
5	(21,21,21,21,21)	100.0139	10.7849	100.3271	11.1283	99.4741	10.7276
10	(22,22,22,22,22)	99.9968	10.5694	100.2684	10.6824	99.4734	10.3374

N=100							
R	**n**	G. Pathak	Std Dev.	M_{fp} MLE	Std Dev.	M_p CUE	Std Dev
0	(25,25,25,25,25)	99.9771	7.9520	99.9827	8.1174	99.5432	8.0327
10	(30,30,30,30,30)	99.9728	6.7360	99.8387	6.7338	99.7129	6.6646

Table 2.2: Expected values from 1000 realisations of the estimators under various M_{fp} scenarios for a population, N=100.

Tables 2.1 and 2.2 select just a few fixed sample size trials and compare the three estimators' means and standard deviations, with and without the inclusion of plants. Some adjustment is made to the sample sizes when plants are included to ensure that the expected number of target animals captured in the trial is comparable to the corresponding no-plant case. The effect of including plants is evident most strongly by comparing the first two trial rows in each table. In Table 2.1 it can be seen that the mean estimates from all three estimators more than treble in size with the inclusion of 10 plants, from an initial estimate that is far too low for each estimator. The generalised Pathak and M_p CUE still have mean estimates that underestimate N, but these means are much improved. The M_{fp} MLE increases from a mean estimate that underestimates N by about 65% when no plants are present to a mean estimate which overestimates N by about 10% when 10 plants are present. It can be seen that including plants mostly reduces the sample standard deviations, but there are some exceptions. In the first two rows of each table the sample standard deviation for all three estimators increases when 10 plants are included. This can be explained, however, by the fact that the mean estimates increase by as much as they do. It is also evident that in the other trials tabulated, the inclusion of plants usually decreases the sample standard deviation, but the magnitude of this difference decreases as the number of captures in each sample increases. The exception to this is the comparison of the last two trials given in Table 2.1, where the sample standard deviation actually increases for all three estimators when plants are included. This may be a result of simulation error due to the fact that the initial sample standard deviation is very low compared to the mean estimate.

It is difficult to determine the optimal estimator based on the sample standard deviations, as no one estimator has a consistently lower sample standard deviation than the others.

When comparing Tables 2.1 and 2.2 it appears that there is a similar pattern in both tables, suggesting that the results are independent of N. This is a desired property, as it is better for the practitioner if the optimal estimator is not a function of the unknown population size, N.

It is concluded here, however, that the generalised Pathak estimator be the proposed estimator under model M_{fp}.

Also given in Table 2.1 is a set of three trials where there is an unequal number of captures in each sample. This is not explored any further in this thesis, but these trials suggest that the M_p CUE may not perform as well when there is an unequal number of animals captured in each sample. This unequal sample size element may replicate heterogeneity amongst the animals, which is not captured in the model M_p conditionally unbiased estimator.

The three trials also prove inconclusive when trying to determine whether adding plants improves estimation. This suggests that further work is required in order to extrapolate the conclusions given above to the unequal sample sizes case.

2.5.2 The optimal number of plants

When adding plants, the two aims are to add as few plants as possible to satisfy the assumptions required more easily but to add as many planted individuals as required to make a worthwhile improvement in estimation. The previous subsection used 10 plants for the trials where there was a non-zero number of plants. That choice is now examined.

Given below is an illustration of how the insertion of different numbers of plants improve the generalised Pathak estimator in the cases when $N = 50$ and $N = 100$. In each case, there are five samples in each trial ($t = 5$). It would be beneficial to the practitioner to get a good population size estimate with as few samples as possible, as this would be less expensive, time-wise and financially, and less invasive for the species being targeted.

With the generalised Pathak estimator it was noted above that the estimator produced almost unbiased mean estimates for most of the range of sample sizes tried. Improvement was needed at just the left-hand end of the plots, where there were few captures in each sample. Thus, for populations $N = 50$ and $N = 100$ the range $0, 1, \ldots, \frac{N}{5}$ captures in each sample is plotted in Figures 2.8 – 2.9 to illustrate the effect of different numbers of plants.

Firstly, for $N = 50$, Figure 2.8 gives the generalised Pathak estimate for the range when there are no captures in each sample to when there are 10 captures in each sample. At the higher end, it is clear that there is no improvement required but the lower end shows that, for each 5 additional plants included, the mean estimated population size is much improved from the previous number of plants. Table 2.3 illustrates the lower end of Figure 2.8, giving the expected value of the generalised Pathak estimator for the lower end of the plot. The largest increase in mean estimate is seen around $\{n_j = 2,\ j = 1, \ldots, t\}$ and $\{n_j = 3,\ j = 1, \ldots, t\}$, where the generalised Pathak mean estimate when $R = 0$ is still poor but the inclusion of plants begins to give a reasonable mean. With $R = 10$ and $R = 15$ the expected value is within 10% of $N = 50$ from $\{n_j = 3,\ j = 1, \ldots, t\}$ onwards. Thus, as fewer plants are deemed more desirable, the recommendation is to use 10 plants, as was done above. However, it is evident that adding only 5 plants can improve the quality of estimation quite considerably.

For $N = 100$, Table 2.4 summarises the left-hand end of Figure 2.9, which gives the expected values of the point population estimate from the generalised Pathak estimator for the trials where there are up to 20 captures in each sample. Again,

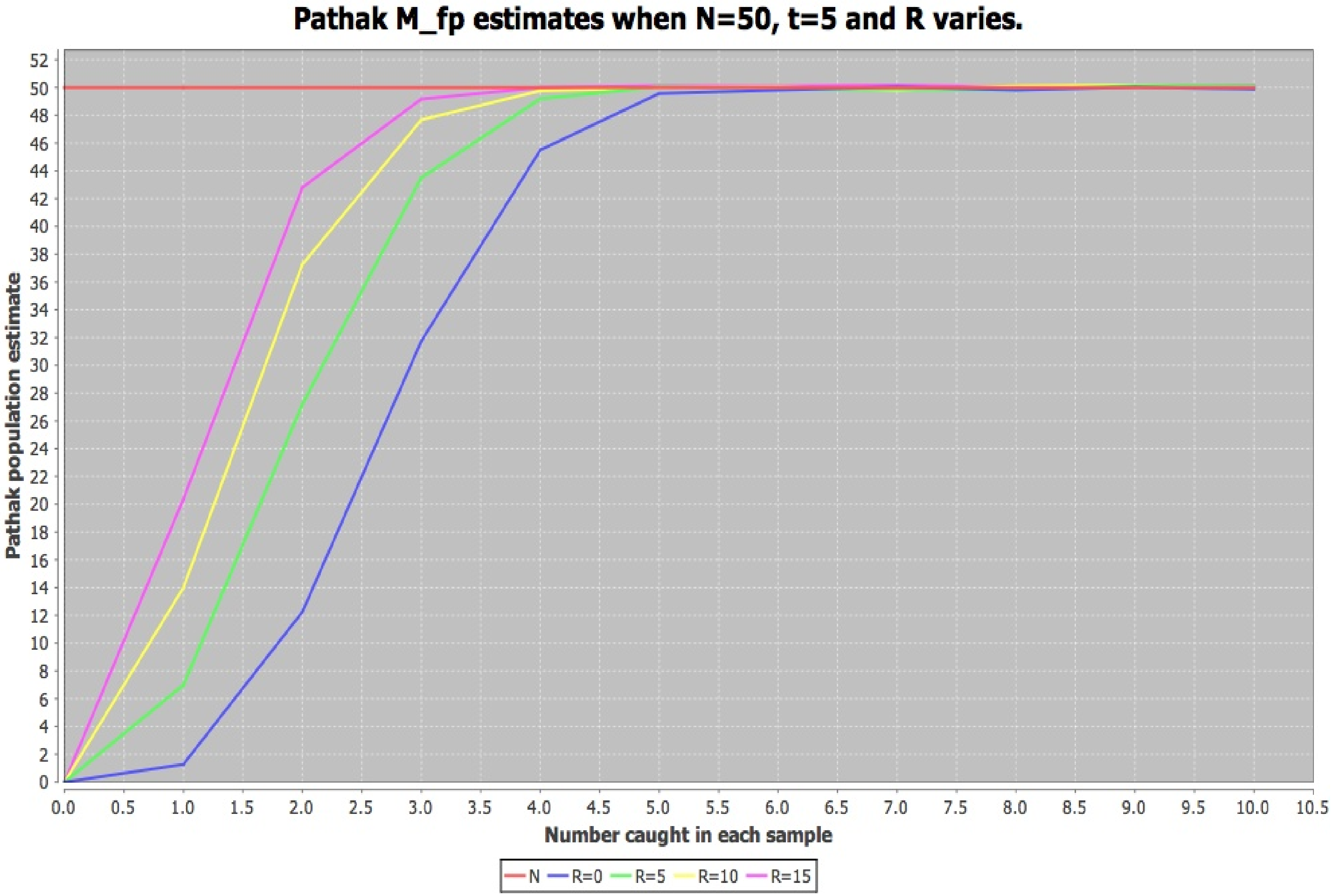


Figure 2.8: Plots of the expected value of the generalised Pathak estimator when $N = 50$ for the equal sample sizes case and varying R.

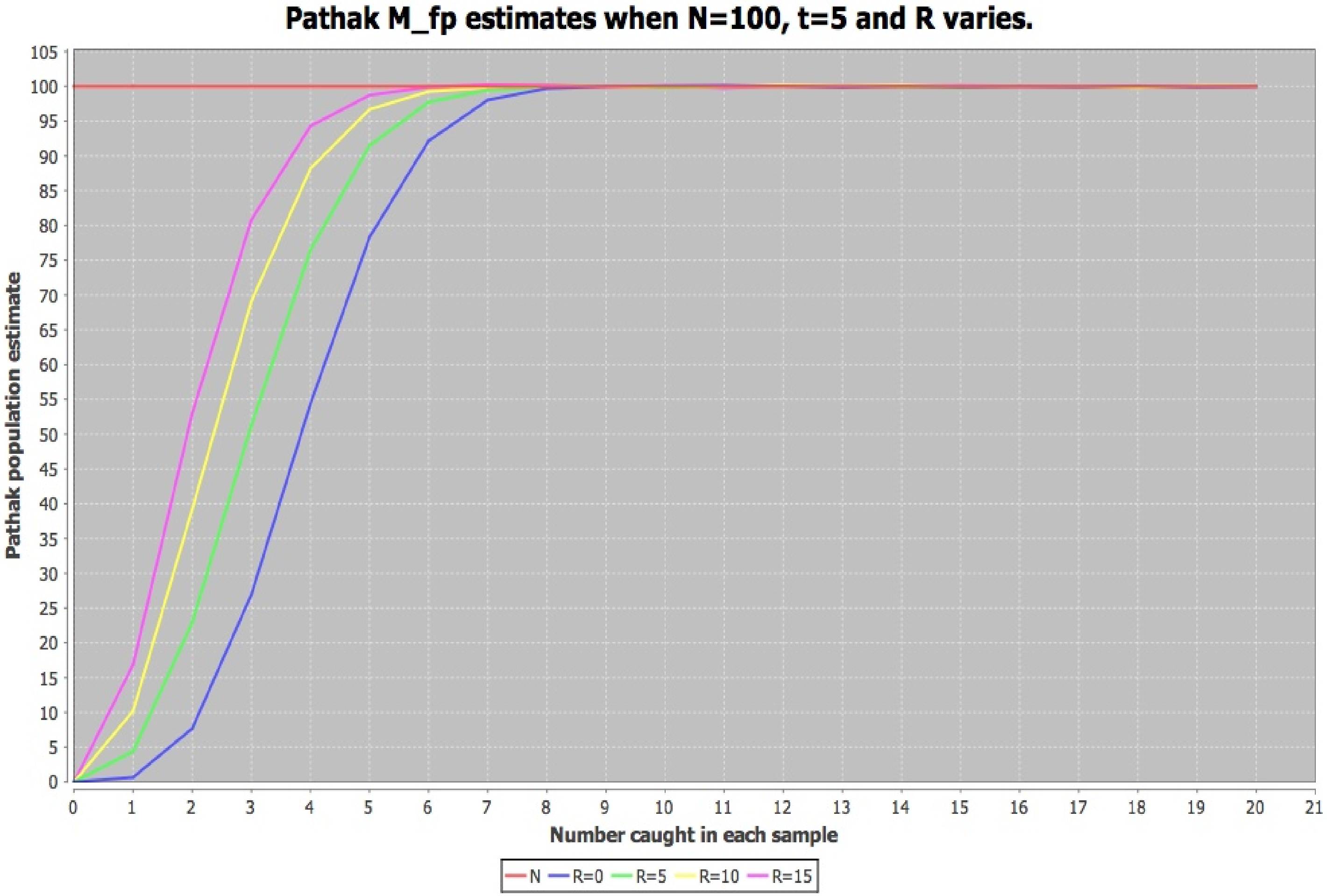


Figure 2.9: Plots of the expected value of the generalised Pathak estimator when $N = 100$ for the equal sample sizes case and varying R.

Number of plants	Number of animals caught in each sample				
	$n_j = 1$	$n_j = 2$	$n_j = 3$	$n_j = 4$	$n_j = 5$
$R = 15$	20.38	42.80	49.16	50.00	50.13
$R = 10$	14.00	37.26	47.67	49.77	50.00
$R = 5$	6.99	27.20	43.50	49.18	50.06
$R = 0$	1.28	12.25	31.74	45.50	49.57

Table 2.3: The expected value of the generalised Pathak estimator when $N = 50$, $t = 5$ and a varying number of plants and sample size.

Number of plants	Number of animals caught in each sample				
	$n_j = 1$	$n_j = 2$	$n_j = 3$	$n_j = 4$	$n_j = 5$
$R = 15$	16.87	52.97	80.79	94.29	98.73
$R = 10$	10.21	39.19	69.09	88.17	96.69
$R = 5$	4.39	22.93	51.28	76.45	91.48
$R = 0$	0.66	7.69	26.93	54.34	78.31

Table 2.4: The expected value of the generalised Pathak estimator when $N = 100$, $t = 5$ and a varying number of plants and sample size.

when $\{n_j = 4,\ j = 1, \ldots, t\}$, adding 15 plants results in an expected value within 10% of N. If an additional animal is caught in each sample, then all three trials with the inclusion of plants have means within 10% of N. The trial where $R = 0$ only has an expected value of 78.31, however. Thus, the case for using plants is quite strong in this situation.

An interesting question posed by this case is whether the addition of 5 more plants improves point population estimation better than the capture of an additional animal in each sample. This question will not be answered in this thesis.

2.5.3 Comparing the Pathak estimator with its approximation

Another analysis carried out here is to compare the Pathak estimator ((2.10) with $R = 0$) with the approximation (2.16) given in Pathak (1964, p. 79). Pathak states in his paper that his estimator (2.10) is difficult to compute, except when $n_1 = n_2 = \ldots = n_t = 1$. The above sections have shown that the generalised Pathak estimator can now be computed for the whole possible range (when the sample sizes are constant) of $n_j \in \{0, x + R\},\ j = 1, \ldots, t$ with accuracy. Computation can still prove troublesome, however, so it is of interest to see how the approximation performs.

As noted in Pathak (1964, p. 79), both estimators (2.10) and (2.16) are equal when $n_1 = n_2 = \ldots = n_t = 1$, as shown in Figures 2.10–2.12. It is evident, however, that the approximation then tends to overestimate the true population size as the number of captures in each sample increases, leading to estimates outside of one standard deviation of the Pathak estimator, (2.10). As the generalised Pathak estimator is almost unbiased for the range of values calculated, a narrow confidence interval was used for the comparison. It can be seen that, as N increases, the range which the Pathak approximation is inside the generalised Pathak confidence interval gets smaller. The Pathak approximation does, however, give estimates in only a few seconds at most for the population and sample sizes in the trials here, whereas the generalised Pathak estimator can take slightly longer as N increases.

Thus, the Pathak approximation estimator, (2.16), cannot be recommended for anything other than offering an upper bound to N, except when the sample sizes are small.

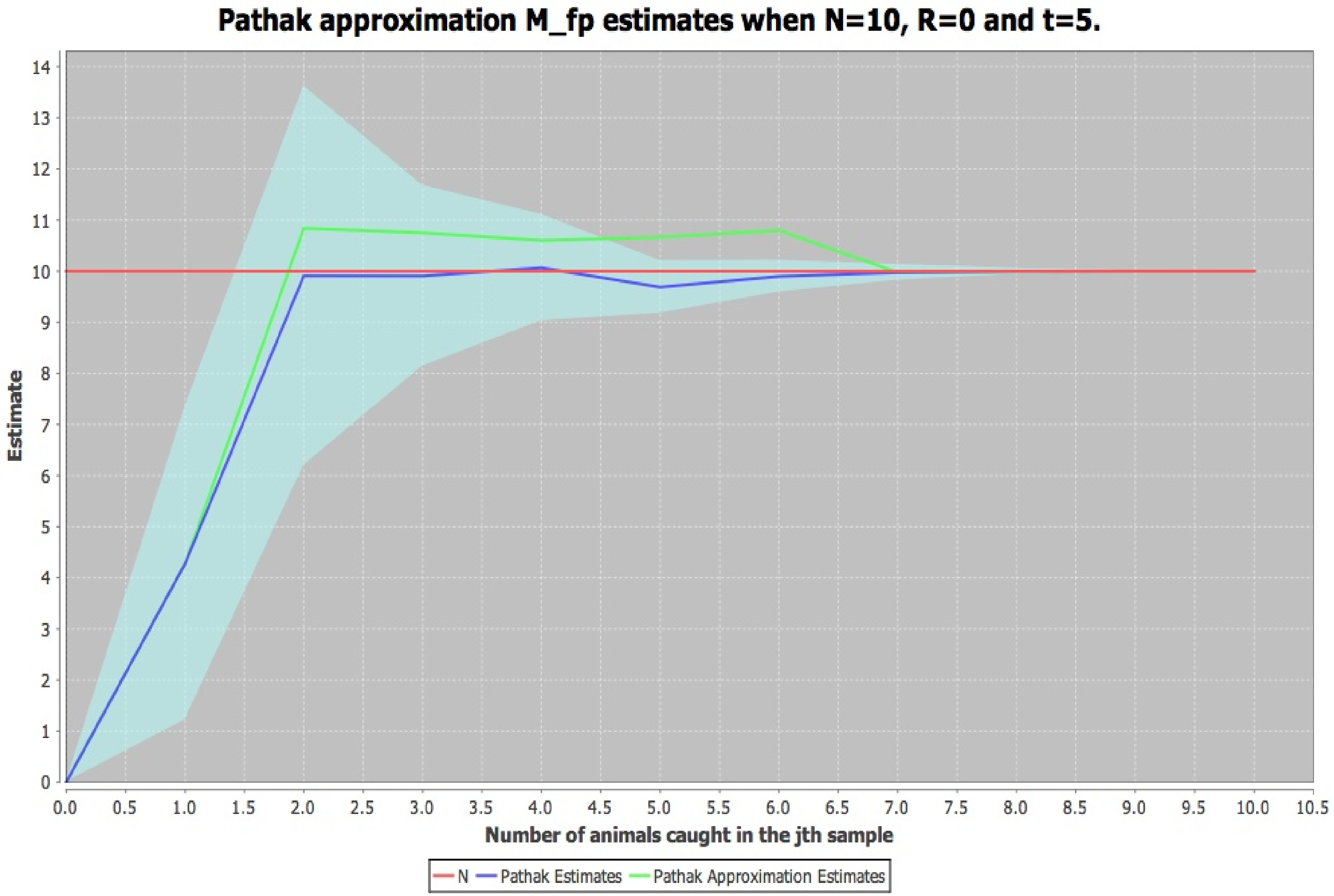


Figure 2.10: Plots of the expected value of the generalised Pathak estimator when $N = 10$ and $R = 0$ with $\pm$ 1 standard deviation of the mean shaded in light blue, along with the Pathak approximation overlaid.

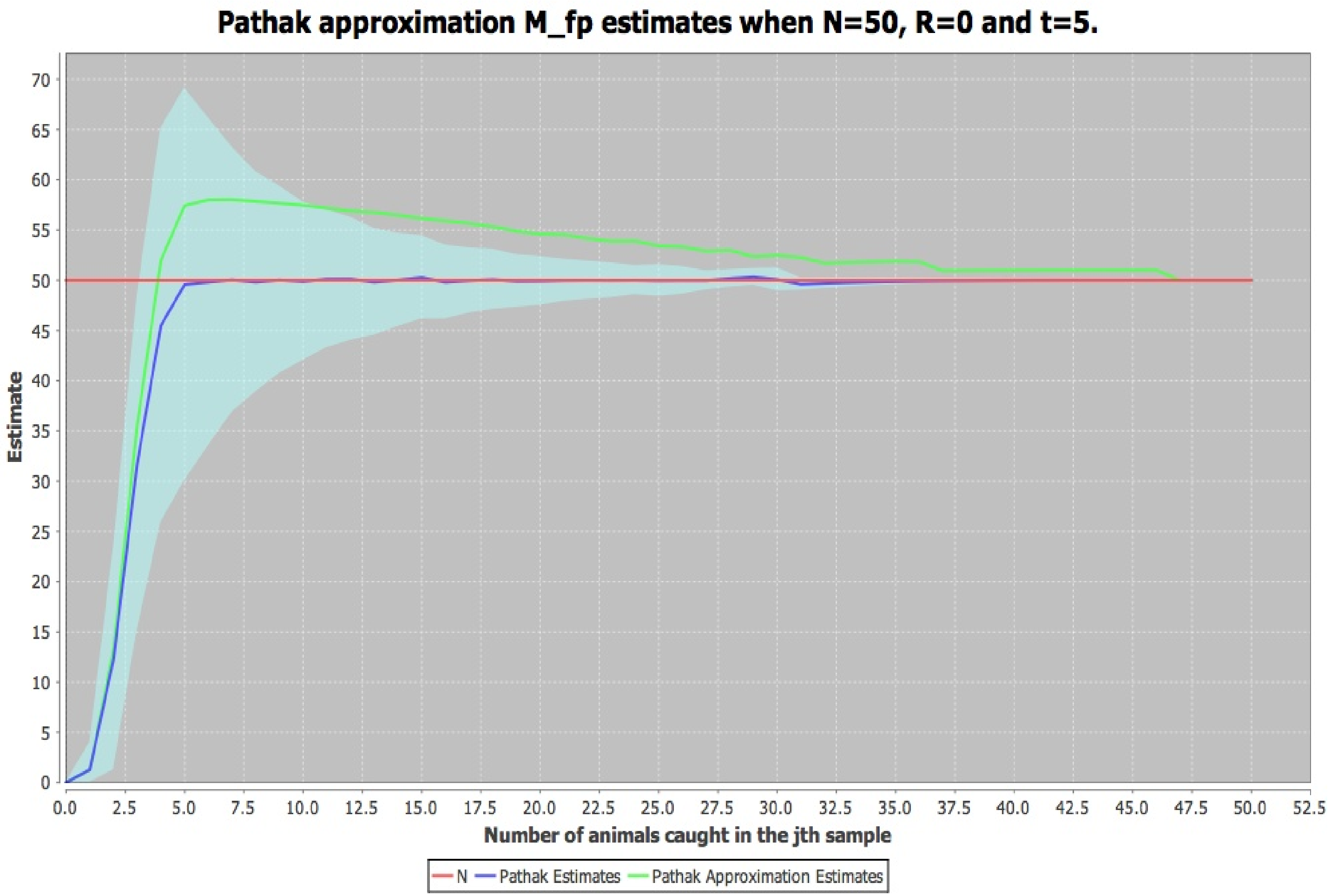


Figure 2.11: Plots of the expected value of the generalised Pathak estimator when $N = 50$ and $R = 0$ with ± 1 standard deviation of the mean shaded in light blue, along with the Pathak approximation overlaid.

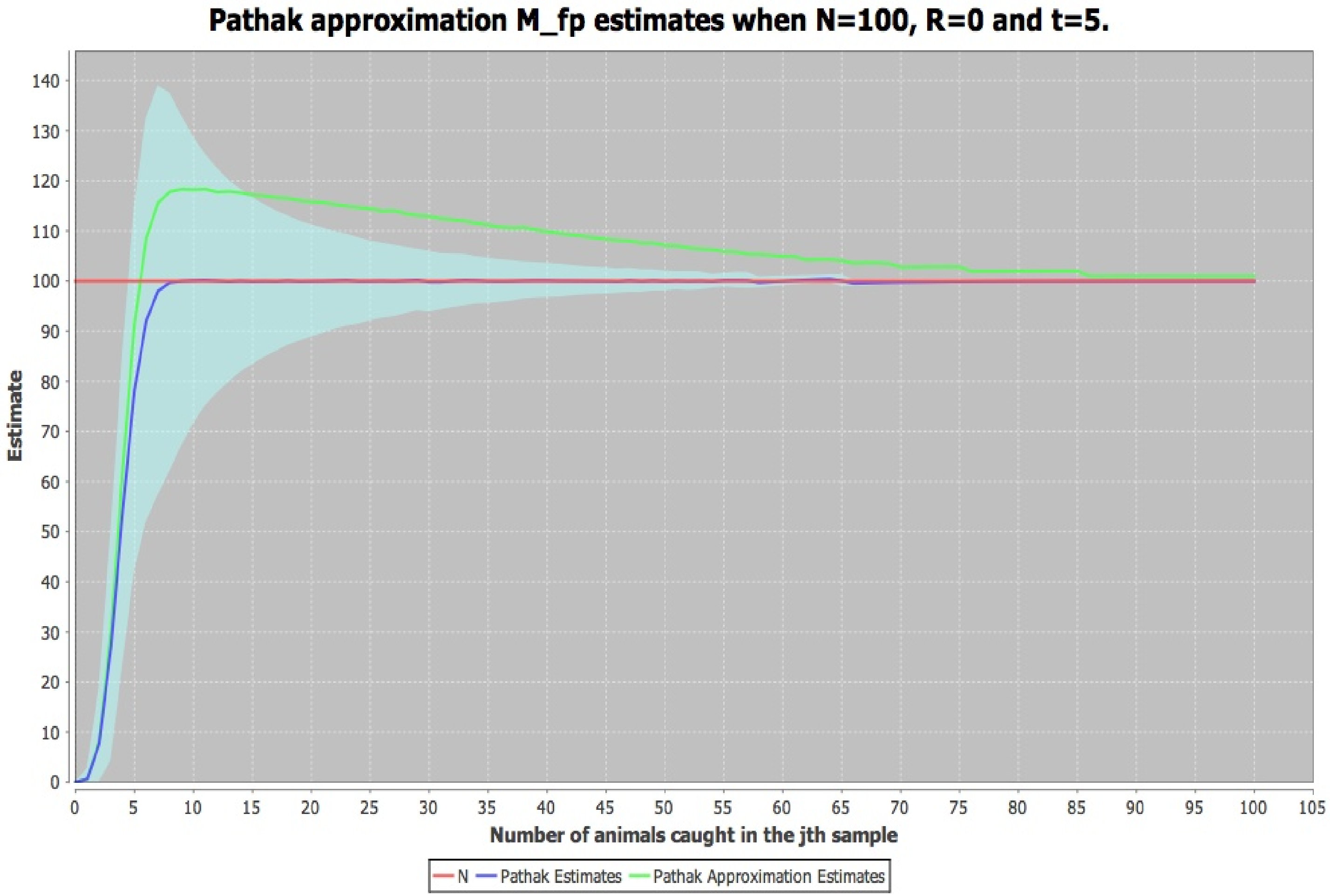


Figure 2.12: Plots of the expected value of the generalised Pathak estimator when $N = 100$ and $R = 0$ with ± 1 standard deviation of the mean shaded in light blue, along with the Pathak approximation overlaid.

2.6 Summary

The strongest conclusion that can be drawn from this chapter is that adding plants to the population before beginning sampling has some benefits. The improvement in estimation that comes from the additional information gained from the plants can be large, and so should be recommended if the effort is not too great and the behavioral assumptions of including plants is met. From Tables 2.3 and 2.4, it can be seen that adding plants can, in cases where few captures are made in each sample, give the generalised Pathak estimator an expected value much greater than the corresponding estimator in a trial without plants present.

The graphs given in §2.5.1 also illustrate that the M_{fp} MLE is prone to overestimating the true population for some values of Z, which is an undesirable quality in ecological circumstances. This overestimation problem is not evident in either the generalised Pathak estimator or the M_p CUE, leading to these estimators being preferred over the M_{fp} MLE under model M_{fp} conditions.

Also shown in this chapter is the justification for using 10 plants, although there is evidence to suggest that $R \geq 5$ should also improve estimation over the no-plants case.

One other point illustrated in this chapter is the poor performance of the Pathak estimator approximation, (2.16), which is guilty of overestimating the true population size in most of the trials modelled here.

Thus, it is the recommendation of this chapter that the generalised Pathak estimator, (2.10) should be used under model M_{fp}.

Chapter 3

ESTIMATION OF POPULATION SIZE UNDER HOMOGENEITY

3.1 Introduction

Of the Otis-class of models (Otis et al. 1978) the homogeneous model, M_0, is the simplest model and has been extensively covered in the literature. A novel aspect will be to apply a Pathak's estimator, which was derived under model M_f, to model M_0. This estimator will be compared to the Pathak estimator Rao-Blackwellised under M_0 to give the conditionally unbiased estimator under M_0.

Model M_p, on which there is less literature, will be covered here also. The generalised Pathak estimator given above will be compared with the conditionally unbiased estimator given by Ashbridge (Ashbridge (1998), Goudie & Ashbridge (2005)), and some results stated to determine whether the inclusion of plants offers improved population estimation.

3.2 Probability Theory

For the probability theory we summarise §1 of Goudie et al. (2007). This chapter mainly deals with model M_p explicitly, but results for the non-plant model M_0 can be derived by letting $R = 0$ wherever it appears. Thus, both mark-recapture and plant-capture distributions are given here.

For finding the probability of capturing z animals, given that the probability of capturing any animal in any sample is p, one naturally chooses a binomial distribution

$$p(Z = z|N, p) \quad = \quad \binom{Nt + Rt}{z} p^z (1-p)^{Nt+Rt-z}, \ z \in \{0, 1, \ldots, Nt + Rt\} \tag{3.1}$$

since each of the $Nt+Rt$ elements of D (c.f. (1.4)) gives the outcome of a Bernoulli trial in which an animal is captured. These probabilities are then scaled to sum to unity for all possible values of $z \in (0, \ldots, Nt + Rt)$. It is then shown that the probability of capturing x distinct animals, given that z are captured in total, is given by

$$p(x|Z = z, N, p) = \frac{(N)_x \, G\left(z, x, t, Rt\right)}{\left(Nt + Rt\right)_z}, \tag{3.2}$$

where $x \in \{\max\{n_j\}, \ldots, \min(N, z)\}$ and $G(z, x, t, Rt)$ is as defined in (2.19). The Gould-Hopper numbers can also be calculated using a triangular recurrence relation, given by

$$G(z+1, x, t, Rt) = (xt + Rt - z)G(z, x, t, Rt) + tG(z, x-1, t, Rt) \qquad \begin{aligned} z &= 0, 1, \ldots; \\ x &= 0, 1, \ldots \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} G(z, 0, t, Rt) &= (Rt)_z \qquad && z = 0, 1, \ldots \\ G(0, x, t, Rt) &= 0 && x = 1, 2, \ldots. \end{aligned} \tag{3.4}$$

When $R = 0$, this Gould-Hopper number simplifies to a C-number, as given by Charalambides & Singh (1988), which is defined as

$$C(z, x, t) = \frac{1}{x!}\sum_{\nu=0}^{x}(-1)^{\nu}\binom{x}{\nu}\Big(t(x-\nu)\Big)_z = \frac{1}{z!}\Delta^x\left[(Nt)_z\right]_{N=0}. \tag{3.5}$$

From (3.1) and (3.2) we get the joint distribution of z and x to be

$$p(z, x|N, p) = \frac{(N)_x}{z!}G(z, x, t, Rt)p^z(1-p)^{Nt+Rt-z}. \tag{3.6}$$

It can be seen from this that, since the unknown parameters (N, p) only appear in terms along with the parameters (z, x), the variables Z and X are the sufficient statistics under models M_0 and M_p.

There is a nice result given by Ashbridge & Goudie (2009, p. 3), analogous to Goudie & Ashbridge (2005, p. 1547) under model M_0, which gives (3.6) as a recurrence relation under model M_p. The joint probability function of z and x, conditional on N and p, satisfies the recursion

$$p(z, x) = \Big[(xt + Rt - z + 1)p(z-1, x) + (N - x + 1)tp(z-1, x-1)\Big]\left[\frac{p}{z(1-p)}\right] \tag{3.7}$$

with starting conditions

$$p(z,0) = \binom{Rt}{z} p^z (1-p)^{Nt+Rt-z} \qquad \text{for } z = 0, \ldots, Rt$$

and

$$p(x,x) = \binom{N}{x} p^x \, (1-p)^{Nt+Rt-x} \qquad x = 1, \ldots, N$$

with

$$p(z,x) = 0 \qquad x > z > 0.$$

This recurrence relation form is useful when the expected values of various estimators are sought, since the computational time is much reduced and reduces the need for simulation. This is shown in the next section.

3.3 Estimators

Various estimators have been tested on M_0 and M_p scenarios. The proposed estimator in this chapter is an estimator generalised from an estimator of Pathak (1964) that is unbiased, with minimum variance, under model M_f, under the condition that the total number of animals captured over all samples exceeds the total population size.

Another estimator that was considered was an estimator derived by Goudie & Ashbridge (2005), that was shown to be conditionally unbiased in the M_0 case when $z \geq N$, and so takes the name *Conditionally Unbiased Estimator*.

The most commonly used estimator under model M_0 is the M_0 MLE, derived by maximising (3.6). This is not a closed-form estimator, since it seeks to compute the optimal value for $\hat{N}$ amongst all permissible values, which come from the set $\mathcal{N} = \{x+1, x+2, \ldots\}$. It also requires the condition that $z > x$ if the estimator is to remain finite, since in the case when $z = x$, the likelihood is monotonically increasing through $\mathcal{N}$, and so $\hat{N}$ is infinite.

The purpose of this chapter is to provide a comparative study between the commonly used estimators and the under-used generalised Pathak estimator, and also to see how the inclusion of plants affects the results in terms of bias and standard deviation.

3.3.1 Generalised Pathak Estimator

It was shown in §2.5 that the generalised Pathak estimator had a lower bias than the M_{fp} maximum likelihood estimator under model M_{fp}. The generalised Pathak estimator also has the benefit of being a closed-form estimator under both model

M_{fp} and M_p, whereas the corresponding MLEs have no such closed form.
As was given in §2.3.2 the generalised Pathak Estimator has the form

$$\tilde{N}(x, \mathbf{n}, R) = x + \frac{a(x-1, \mathbf{n}, R)}{a(x, \mathbf{n}, R)}, \tag{3.8}$$

where $a(x, \mathbf{n}, R)$ is as given in (2.4).
Under model M_p $\mathbf{n}$ is a random sample with each animal having the same probability of capture over all samples.

3.3.2 M_p Conditionally Unbiased Estimator

The generalised Pathak estimator was shown to be the MVU estimator under model M_{fp}, where there is only one sufficient statistic, X. As shown above, in model M_p there are two sufficient statistcs, X and Z. The generalised Pathak estimate (3.8) is here Rao-Blackwellised under M_p, giving the conditionally unbiased estimator under model M_p (c.f. (Ashbridge & Goudie 2009)). The Rao-Blackwell theorem is stated first:

Rao-Blackwell Theorem: Let $\hat{\theta}$ be an estimator of θ with $E[\hat{\theta}^2] < \infty$. Now let T be sufficient for θ and $\theta^* = E[\theta|T]$. Then

$$E[\theta^* - \theta]^2 \leq E[\hat{\theta} - \theta]^2 \qquad \forall\, \theta.$$

■

The generalised Pathak estimator is Rao-Blackwellised over the set $\{\mathbf{n}_{\{z,x\}} | n_1 + \ldots + n_t = z, n_i \leq x \text{ for } i = 1, \ldots, t\}$, where the n_js here are assumed to be random:

$$\begin{aligned}
E[N^*|Z, X] &= x + \sum_{\mathbf{n} \in \mathbf{n}_{z,x}} \frac{a(x-1, \mathbf{n}, R)}{a(x, \mathbf{n}, R)} \frac{z! a(x, \mathbf{n}, R)}{G(z, x, t, Rt)} \\
= x &+ \frac{z!}{G(z, x, t, Rt)} \sum_{\mathbf{n} \in \mathbf{n}_{z,x}} \frac{1}{(x-1)!} \sum_{k=0}^{x-1} (-1)^k \binom{x-1}{k} \prod_{j=1}^{t} \binom{R + x - 1 - k}{n_j} \\
= x &+ \frac{z!}{(x-1)! G(z, x, t, Rt)} \sum_{k=0}^{x-1} (-1)^k \binom{x-1}{k} \sum_{\mathbf{n} \in \mathbf{n}_{z,x}} \prod_{j=1}^{t} \binom{R + x - 1 - k}{n_j} \\
= x &+ \frac{z!}{(x-1)! G(z, x, t, Rt)} \sum_{k=0}^{x-1} (-1)^k \binom{x-1}{k} \binom{t(R + x - 1 - k)}{z} \\
= x &+ \frac{G(z, x-1, t, Rt)}{G(z, x, t, Rt)}.
\end{aligned}$$

Thus, we have

$$\tilde{N}_U(z,x) = \begin{cases} x + \left[\dfrac{G(z,x-1,t,Rt)}{G(z,x,t,Rt)}\right] & z = 1,2,\ldots; \quad x = 1,\ldots,z; \\ 0 & x = 0;\ z = 0,1,\ldots,Rt. \end{cases} \tag{3.9}$$

This estimator is hereby referred to as the model M_p Conditionally Unbiased Estimator, or M_p CUE, proposed by Ashbridge & Goudie (2009). The estimator was shown in Goudie & Ashbridge (2005) to be unbiased under M_0, under the condition that the total number of captures, z, exceeds N. It has the benefit of being a closed-form estimator, with a probability distribution that can be evaluated via a recurrence relation or by closed-form calculation. Thus, there are two possible methods that can be employed in calculating M_p CUE estimates, namely analytical and via simulation approximation. Exact computation is achieved by multiplying $\tilde{N}_U(z,x)$ with the corresponding $p(z,x)$, as given in (3.6), for each $Z = z$ and $X = x$ and summing over all (z,x).
Alternatively, simulation can be used for calculating the moments, using the recursive estimate $\tilde{N}_U$ found by rearranging

$$\frac{\tilde{N}_U(z,x) - x}{\tilde{N}_U(z-1,x) - x} = \frac{t\tilde{N}_U(z-1,x-1) + Rt - z + 1}{t\tilde{N}_U(z-1,x) + Rt - z + 1}. \tag{3.10}$$

In order to start the recursion some boundary conditions are required. These can be determined from the Gould-Hopper numbers triangular recurrence relation, (3.3), with boundary conditions given by (3.4), giving:

$$\begin{aligned} \tilde{N}_U(z,x) &= \begin{cases} x & x = 1,\ldots,N;\ z = (R+x-1)t+1,\ldots,(R+x)t \\ 0 & x = 0;\ z = 0,\ldots,Rt \end{cases} \\ \tilde{N}_U(z,z) &= \frac{z(2Rt + tz - z + t + 1)}{2t} \quad \text{for} \quad z = 0,1,\ldots. \end{aligned}$$

Following Goudie & Ashbridge (2005), what was actually used in the simulation was $\hat{N}_U$ where $\hat{N}_U = [\tilde{N}_U + 0.5]$ and the square brackets denote the integer part. This is done to provide a fairer comparison with the M_p MLE, as the MLE can only be integer valued, so this condition is also added to the non-MLE estimators.

3.3.3 M_p Maximum Likelihood Estimator

As with model M_{fp} (c.f. p.2.3.1) there is not a closed-form estimator for the M_p MLE and it also has the condition that z must exceed x in order for the maximum to remain finite. Thus, we seek the maximum for the likelihood given in §1.6.2. This is done by differentiating (3.6) with respect to p and setting equal to zero. Hence, we find that p is maximised when

$$\hat{p} = \frac{z}{t(N+R)}.$$

Thus, similar to what Otis et al. (1978, p. 105) did under model M_0, we substitute this into (3.6) to get the M_p MLE $\hat{N}$ of N to be solution of

$$\begin{aligned}\ell(\hat{N},\hat{p}|X) = \max_{N\epsilon\mathcal{N}}\Bigg[&\ln\left(\frac{N!}{(N-x)!}\right) + z\ln(z) \\ &+ (t(N+R)-z)\ln(t(N+R)-z) - t(N+R)\ln(t(N+R))\Bigg], \quad (3.11)\end{aligned}$$

This requires sequential calculations through the range of $\mathcal{N}$ to find the maximum. This is unhelpful, but Goudie et al. (2007), showed that, for M_p, the profile likelihood is unimodal, so iterations can stop once a turning point has been reached. This goes some way to shortening the computational time involved.

3.4 M_p Computation

3.4.1 Methods

Under M_p, (N, p) is sufficient, and the M_p CUE and M_p MLE are functions of Z and X. Thus, it is relatively straightforward to produce exact moments for these estimators. As the generalised Pathak estimator is a function of X and $\mathbf{n}$, however, producing exact moments is rather more difficult. Thus, in order to compare all three estimators, simulation is the chosen method.

For each trial given below, 1000 realisations from randomly generated capture histories for a specified set of parameters N, p, R and t are tabulated. The estimator in widespread use for model M_0 is the M_0 MLE. Since this requires the condition $z > x$, then it was deemed appropriate initially to attach this condition to all estimators, despite the closed form estimators remaining finite without such a condition. This allows for a fairer comparison between the estimators. However, the unconditioned generalised Pathak estimator and the M_p CUE are also given in Tables 3.1 and 3.2.

The M_0 and M_p cases assume that $p_j = p$, a constant, *ie*, the probability differs neither between animals nor between samples. Simulation results for three values of p are given in the Tables below, namely $p = 0.05, 0.1$ and 0.2.

3.4.2 Results

Some general results can be stated first. Applying the condition that z be greater than x on the closed form estimators lowers their mean population estimate in most cases when $R = 0$ and $t = 5$. When estimation under M_0 is poor, the performance of the estimators in terms of mean estimate is improved by the inclusion of plants. This has the knock-on effect of increasing the sample standard deviation. This is to be expected since the range of individual estimates has increased. The estimators are bounded below by x, and so a very negatively biased estimator has only a small range of possible estimates. By reducing the bias of the estimator, the range of possible estimates will increase.

Also, it can be seen that the CUE and generalised Pathak estimators track each other closely, both in terms of mean estimate and sample standard deviation. Also, the M_0 MLE almost always has the highest mean, which results in it also having the largest standard deviation in almost every trial simulated.

The results for the non-plant homogeneous case, model M_0, are given in Table 3.1. The M_p MLE and the conditional and unconditional forms of both the CUE and Pathak estimator all show a large negative bias when p is small, *ie*, equal to 0.05. This is rectified to some extent by increasing t.

When $N = 10$ there is evidence that 10 trials are preferable to 5 unless the capture probability $p = 0.2$. The exception to this rule is that, when $p = 0.2$, the M_0 MLE mean estimate decreases when t is increased from 5 to 10. This decrease, however, could possibly just be due to simulation error.

When $N = 50$, increasing t from 5 to 10 improves the means of the estimators when $p = 0.05$ quite significantly. However, for $p = 0.1$ or $p = 0.2$, the mean estimates when $t = 5$ are already satisfactory in many cases. Increasing t in these cases, however, does still reduce the sample standard deviations by more than 50%. Thus, there is still a benefit to having more samples, should it be practical to the practitioner.

The results for the heterogeneous plant-capture case, model M_p, are given in Table 3.2. The first observation is that applying the condition that $z > x$ does not affect the generalised Pathak estimator and CUE as much as when $R = 0$. This is as would be expected, as any planted animals that are caught are contained in Z but not X. Thus, the more plants that are included, the smaller the probability of having $z = x$. Also, when comparing Table 3.2 with Table 3.1, it is evident that including plants improves estimation in terms of mean estimate and, in most cases, in terms of the sample standard deviation. The improvement in the mean estimation is most significant when p is small, and the most significant reduction in sample standard deviation occurs when p is high. When $p = 0.05$ the means of the estimators

N	p	Estimator	$t = 5$		$t = 10$	
			Mean	Sample std dev.	Mean	Sample std dev.
10	0.05	Cond. Pathak	3.92	2.30	6.79	3.28
		Uncond. Pathak	4.03	3.21	7.26	4.46
		Cond. CUE	3.93	2.28	6.79	3.26
		Uncond. CUE	4.03	3.15	7.26	4.43
		M_p MLE	4.47	3.29	8.19	5.11
	0.1	Cond. Pathak	6.60	3.21	9.44	3.06
		Uncond. Pathak	7.25	4.29	9.90	3.75
		Cond. CUE	6.61	3.20	9.43	3.02
		Uncond. CUE	7.23	4.19	9.88	3.69
		M_p MLE	7.94	4.91	10.26	4.59
	0.2	Cond. Pathak	9.23	2.94	9.94	1.49
		Uncond. Pathak	9.91	4.11	9.94	1.33
		Cond. CUE	9.25	2.96	9.93	1.48
		Uncond. CUE	9.93	4.09	9.94	1.32
		M_p MLE	10.10	4.51	9.59	1.70
50	0.05	Cond. Pathak	29.31	13.22	48.64	17.86
		Uncond. Pathak	38.85	22.33	48.31	20.58
		Cond. CUE	29.28	13.11	48.59	17.75
		Uncond. CUE	38.85	22.18	48.28	20.54
		M_p MLE	42.74	24.28	58.14	31.05
	0.1	Cond. Pathak	47.46	16.93	50.23	8.47
		Uncond. Pathak	49.56	22.25	49.89	8.42
		Cond. CUE	47.49	16.88	50.24	8.47
		Uncond. CUE	49.56	22.17	49.89	8.43
		M_p MLE	57.21	30.07	50.93	9.25
	0.2	Cond. Pathak	49.61	8.31	49.90	2.94
		Uncond. Pathak	49.58	8.07	49.95	3.07
		Cond. CUE	49.60	8.26	49.50	2.94
		Uncond. CUE	49.59	8.06	49.95	3.07
		M_p MLE	50.24	8.99	49.48	2.98

Table 3.1: Mean estimates under model M_0 of population size based on 1000 bootstrap samples where the condition is whether $z = x$ is permitted or not.

N	p	Estimator	$t = 5$		$t = 10$	
			Mean	Sample std dev.	Mean	Sample std dev.
10	0.05	Cond. Pathak	8.55	6.19	10.02	5.30
		Uncond. Pathak	9.41	8.40	10.16	5.40
		Cond. CUE	8.55	6.19	10.02	5.30
		Uncond. CUE	9.39	8.38	10.15	5.39
		M_p MLE	11.70	10.68	11.09	7.61
	0.1	Cond. Pathak	9.89	5.31	9.93	2.95
		Uncond. Pathak	10.13	5.38	10.02	2.86
		Cond. CUE	9.89	5.32	9.92	2.94
		Uncond. CUE	10.12	5.37	10.02	2.86
		M_p MLE	11.03	7.84	9.74	3.23
	0.2	Cond. Pathak	10.03	2.93	10.03	1.27
		Uncond. Pathak	9.86	2.77	10.04	1.22
		Cond. CUE	10.03	2.93	10.03	1.26
		Uncond. CUE	9.86	2.77	10.04	1.22
		M_p MLE	9.83	3.22	9.53	1.34
50	0.05	Cond. Pathak	46.81	22.86	50.30	15.38
		Uncond. Pathak	48.92	31.55	49.14	15.66
		Cond. CUE	46.82	22.86	50.31	15.38
		Uncond. CUE	48.91	31.59	49.14	15.65
		M_p MLE	63.03	43.61	53.87	19.47
	0.1	Cond. Pathak	50.00	16.63	49.73	7.65
		Uncond. Pathak	50.39	15.83	50.25	7.50
		Cond. CUE	49.98	16.61	49.73	7.64
		Uncond. CUE	50.40	15.83	50.25	7.50
		M_p MLE	53.74	22.62	49.93	7.97
	0.2	Cond. Pathak	50.27	7.23	49.78	2.87
		Uncond. Pathak	49.82	6.93	50.03	2.84
		Cond. CUE	50.28	7.22	49.78	2.87
		Uncond. CUE	49.83	6.93	50.03	2.83
		M_p MLE	50.44	7.52	49.36	2.89

Table 3.2: Mean estimates under model M_p of population size, augmented by 10 plants, based on 1000 bootstrap samples where the condition is whether $z = x$ is permitted or not.

under M_0 are shown to be quite poor, underestimating N by more than 50% when $N = 10$. The mean estimates under M_p when $p = 0.05$ are much closer to N.

3.5 Conclusion

It is shown in Tables 3.1 and 3.2 that the M_p CUE and generalised Pathak estimator produce very similar estimates for the cases simulated. Due to the nature of simulation, one cannot analyse the numbers in fine detail and must be more cautious when drawing conclusions. However, it seems evident that both the generalised Pathak estimator and the M_p CUE have very similar means and sample standard deviations for the cases simulated here. The means from both these estimators are shown to be superior to that of the M_0 and M_p MLEs in many of the cases simulated. The M_0 MLE is evidently better when $p = 0.05$ and $t = 5$, the case where all estimators are negatively biased, but increasing t to 10 results in a mean overestimation when $N = 50$. The inclusion of plants can equally be seen to result in the MLE overestimating when $p = 0.05$. Thus, the M_p MLE is not uniformly better than the M_0 MLE.

In almost all trials, the MLE has a larger standard deviation than the other estimators. This seems intuitive, as the CUE is a Rao-Blackwellised estimator under models M_0 and M_p. The generalised Pathak estimator is the MVUE under models M_f and M_{fp}, and the results suggest that it also has a very low standard deviation under these models.

There is evidence from the comparison of Tables 3.1 and 3.2 that including 10 plants benefits the generalised Pathak estimator and the CUE. Thus, model M_p is concluded to be more favourable than model M_0, and the generalised Pathak estimator or the M_p CUE recommended for this model.

Chapter 4

Estimation of Population Size under Time-Dependent Capture Probabilities

4.1 Introduction

The main focus of this chapter is to extend the work that has been carried out on plant-capture population estimation under homogeneous models to time heterogeneous models. Model M_{tp} assumes that there is homogeneity between all the animals in any particular sample, but the probability of capture differs between samples. This type of estimation has been carried out several times in recent years, most notably in the 1990 US Census Bureau decennial census. This census aimed to incorporate an estimate of the number of homeless people in the United States (see Laska, Meisner & Siegel (1988), Laska & Meisner (1993) and Martin et al. (1997) for more information on the survey).

The technique has been refined and used in the annual Homeless Outreach Population Estimate (HOPE) survey (Hopper et al. (2008)). This paper details some of the difficulties experienced when inserting plants, and illustrates the importance, under the assumption of capture homogeneity between all individuals, of plants behaving exactly like their native co-habitants.

The non-plant model, M_t, has had a great volume of literature written about it (see Buckland et al. (2000, p. 2), Lin & Chao (2005, pp. 94-96) and references therein), but one of the main goals here is explore the effect that including plants has on the quality of estimation.

One aim of this chapter is to examine the benefits of model M_{tp} over model M_t. Another aim is to examine the performance of the estimator of Pathak (1964) when, unlike the context for which it was designed, the sample sizes are random. Also of interest is to establish whether the computational difficulties of Pathak's estimator, that existed in the era of its proposal, are manageable with the computing power available nowadays.

4.2 Probability Theory

Results for model M_t can be derived from those for model M_{tp} by letting $R = 0$. Model M_{tp} has $t + 1$ parameters, namely $N, p_1, p_2, \ldots, p_t$, where p_j represents the capture probability for all N animals for the j^{th} sample, $j = 1, \ldots, t$. Under this model the number of animals caught in each sample, n_j, $j = 1, \ldots, t$, are independent random variables. Darroch (1958) showed that the probability density for this model is multinomial with parameters $N, p_1, \ldots, p_t$. Generalising this to model M_{tp}, we get

$$p(x, \mathbf{n}|N, \mathbf{p}) = (N)_x \, a(x, \mathbf{n}, R) \prod_{j=1}^{t} p_j^{n_j} (1 - p_j)^{N+R-n_j} \tag{4.1}$$

for $x = 0, \ldots, N$ and $n_j = 0, \ldots, x, \;\; (j = 1, \ldots, t)$, with $a(x, \mathbf{n}, R)$ as given in (2.4).

From (4.1) it can be seen that $(x, \mathbf{n})$ is sufficient for $(N, \mathbf{p})$. It is clear from (4.1) that, since the ranges of x and $\mathbf{n}$ increase with N, this probability space quickly becomes very large, with the capture matrix D sparsely populated with 1s, making exact computation of the properties of estimators difficult.

Taking the logarithm of (4.1) gives the multinomial log-likelihood of $(N, \mathbf{p})$ as

$$\begin{aligned} \ell(N, p_1, \ldots, p_t; x, n_1, \ldots, n_t) &= \ln\left(\frac{N!}{(N-x)!}\right) + \sum_{j=1}^{t} n_j \ln(p_j) \\ &+ \sum_{j=1}^{t} (N + R - n_j) \ln(1 - p_j) + const. \end{aligned} \tag{4.2}$$

This will be used below to calculate the MLE for model M_{tp}.

4.3 Bayesian mark-recapture under model M_{tp}

4.3.1 Introduction

A detailed analysis of Bayesian statistics is not carried out in this thesis. Instead, the reader is directed to the growing list of Bayesian literature, amongst which are McCarthy (2007), King et al. (2009) and Link & Barker (2009), books aimed at ecologists and that assume only knowledge of classical statistical methods.

Under model M_{tp} classical statistical methods state that if the model has parameters $(N, p_1, \ldots, p_t)$, of which inference is made, the data $\mathbf{x} = (x, n_1, \ldots, n_t)$ is collected and used to estimate the fixed parameters. This can be done using the log-likelihood function, (4.2).

The key distinction with Bayesian statistics is that it does not assume that the parameters fixed, but rather they have an underlying distribution that is assigned by the user. This assigned distribution is referred to as the *prior distribution*, denoted by $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters. By multiplying this prior distribution with the likelihood function, denoted here by $f(\mathbf{x}|\boldsymbol{\theta})$, one can get a *posterior distribution* of the parameters, given the data $\boldsymbol{x} = (x_1, \ldots, x_n)$, as

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}). \tag{4.3}$$

Equality can be achieved here by calculation of the normalising constant, but this can be arduous. Calculating this normalising constant is not required for estimation, however, when one uses Markov Chain Monte Carlo (MCMC) methods. As (4.3) gives the posterior distribution of the parameters, it is simple to get summary statistics and construct various intervals. The summary statistics most often used are the mean, median and mode, where each has a theoretical justification for being preferred. Similarly, a $100(1-\alpha)\%$ interval (called a **credible interval**) can be constructed in different ways. A $100(1-\alpha)\%$ credible interval for θ, in the one-dimensional case, is defined as the interval $[a, b]$ that satisfies

$$P(\theta \in [a, b]) = \int_a^b \pi(\theta|\mathbf{x}) d\theta = 1 - \alpha, \qquad 0 \leq \alpha \leq 1. \tag{4.4}$$

It is evident, however, that there are many possible intervals $[a, b]$ that can satisfy (4.4), and so a further definition follows. A **highest posterior density interval**, *HPDI*, is the $100(1-\alpha)\%$ interval $[a, b]$, centred around the mode, satisfying:

1. $[a, b]$ is a $100(1-\alpha)\%$ credible interval;
2. for all $\theta' \in [a, b]$ and $\theta'' \notin [a, b]$, $\pi(\theta'|\mathbf{x} \geq \pi(\theta''|\mathbf{x})$.

This can be generalised to the multi-parameter case quite simply. In many 'nice' cases, the posterior distribution, (4.3), will be a standard statistical distribution, whose moments can be calculated exactly. Not all prior distributions or likelihoods lead to these 'nice' posteriors, however, and in these cases, some other method is required to calculate the moments. This is possible through the use of Markov chain Monte Carlo, MCMC.

One method for estimating the moments is the Gibbs Sampler. A thorough description of its usage is given in Casella & George (1992). From above, the model has parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ with prior distribution, $\pi(\boldsymbol{\theta})$. From this, a set of conditional probabilities, $\pi(\theta_j|\boldsymbol{\theta_{(j)}})$ are drawn, where the vector $\boldsymbol{\theta_{(j)}} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$, with the relevant adjustment for $j = 1$ or $j = k$. A starting point for the Gibbs sequence is specified, say a vector of parameters $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_k^0)$. Then, from these initial conditions, rejection sampling can be used to draw samples from the conditional distributions. To start the iterative step, sampling is as follows:

$$
\begin{array}{lll}
\theta_1^1 & \text{is sampled from} & \pi(\theta_1|\boldsymbol{\theta}_{(1)}^0) \\
\theta_2^1 & \text{is sampled from} & \pi(\theta_2|\theta_1^1, \theta_2^0, \ldots, \theta_k^0) \\
 & \vdots & \\
\theta_j^1 & \text{is sampled from} & \pi(\theta_j|\theta_1^1, \theta_2^1, \ldots, \theta_{j-1}^1, \theta_{j+1}^0, \theta_k^0) \\
 & \vdots & \\
\theta_k^1 & \text{is sampled from} & \pi(\theta_k|\boldsymbol{\theta}_{(k)}^1)
\end{array}
$$

Doing this, $\boldsymbol{\theta}^1$ is generated from $\boldsymbol{\theta}^0$. Continuing this, T vectors can be generated, which, after a burn-in period, should represent random samples from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$. King & Brooks (2008) used $T = 1,000,000$ in their computations, with a burn-in period of 100,000.

4.3.2 Bayesian plant-capture probability theory

This section is an extension of George & Robert (1992) to Bayesian plant-capture scenarios, and has an analogous structure. One can get the results given in George & Robert (1992) by setting $R = 0$. Following their notation, let $\mathcal{D} = (x, \mathbf{n})$ be the collected set of data from a mark-recapture experiment. From (4.1)

$$L(N, \mathbf{p}|\mathcal{D}) \propto \frac{N!}{(N-x)!} \prod_{j=1}^{t} p_j^{n_j} (1-p_j)^{N+R-n_j}. \tag{4.5}$$

Using the fact that N and $\mathbf{p}$ are *a priori* independent, the prior distribution has the form $\pi(N, \mathbf{p}) = \pi(N)\pi(\mathbf{p})$. This independence gives posterior conditional distributions of N and $\mathbf{p}$ to be

$$\pi(N|\mathbf{p},\mathcal{D}) \propto \frac{N!}{(N-x)!}\left\{\prod_{j=1}^{t}(1-p_j)\right\}^N \pi(N) \tag{4.6}$$

$$\pi(\mathbf{p}|N,\mathcal{D}) \propto \left\{\prod_{j=1}^{t} p_j^{n_j}(1-p_j)^{N+R-n_j}\right\}\pi(\mathbf{p}). \tag{4.7}$$

Note that (4.6) is unchanged from that given in George & Robert (1992).

Now let the p_js be *a priori* independent, Be(α,β)-distributed random variables, with mean $\mu = \dfrac{\alpha}{\alpha+\beta}$ and variance $\sigma^2 = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Thus, $\pi(\mathbf{p}) = \prod_{j=1}^{t}\pi(p_j)$, where $\pi(p_j) \sim$ Be(α,β). This gives a posterior conditional distribution for $p_j,\ j = 1,\ldots,t$, as

$$\pi(p_j|N,\mathcal{D}) \sim \text{Be}(n_j+\alpha, N+R+\beta-n_j). \tag{4.8}$$

Combining (4.6) and (4.8) gives

$$\pi(N|\mathcal{D}) \propto \frac{N!}{(N-x)!}\left\{\prod_{j=1}^{t}\frac{\Gamma(N+R+\beta-n_j)}{\Gamma(N+R+\alpha+\beta)}\right\}\pi(N). \tag{4.9}$$

Taking a ratio of successive terms gives

$$\frac{\pi(N+1|\mathcal{D})}{\pi(N|\mathcal{D})} = \frac{N+1}{N+1-x}\left\{\prod_{j=1}^{t}\frac{(N+R+\beta-n_j)}{(N+R+\alpha+\beta)}\right\}\frac{\pi(N+1)}{\pi(N)}. \tag{4.10}$$

In carrying out rejection sampling estimates drawn from (4.9), and calculating the mean of a sufficiently large number of such estimates, a point estimate of the population size N can be found.

4.4 Estimators

The intention here is to compare the performance of a quintet of estimators: the M_{tp} MLE, the generalised Pathak estimator, the M_p MLE, the M_p CUE and a modified Petersen estimator. These were simulated both with and without plants, and various summary statistics noted to determine optimality. The M_p estimators were initially included to be a measure of the time-heterogeneous effect of the data. Following the results of Chapter 2, however, it is the intention to test whether they can estimate the true population size to a similar level of accuracy to the M_{tp} estimators.

The Petersen-type estimator is generalised to the multiple recapture case, utilising the inclusion of plants. The Petersen estimator lends itself naturally to plant-capture scenarios, where it can be assumed that the ratio of the number of distinct plants captured to those inserted would be expected to be similar to the corresponding ratio of numbers caught from the target population. When this estimator was used in non-plant simulations, it uses the first sample of any simulation to be a 'planting' occasion, and samples $2, \ldots, t$ in standard mark-recapture fashion. It would then mimic a Petersen-type estimator, which is detailed below. As the Petersen estimator has a natural stratification, it may be that this estimator performs better under model M_{tp} than under model M_t.

4.4.1 Maximum Likelihood Estimator for M_{tp}

The most commonly used estimator under model M_t is the model M_t MLE. In the case of no plants, the model M_t MLE, as given in Darroch (1958), Otis et al. (1978) or Seber (1982) using various proofs, can be derived from (4.2) by setting $R = 0$. This estimator is employed by various computer packages, including *MARK*. This MLE is generalised here to accommodate for the inclusion of plants, using the log-likelihood function (4.2). Two shortcomings of the MLE are that it is not, for $t > 2$, a closed-form estimator and that it is infinite when $z = x$.

The MLE is given as the value of N that maximises the log-likelihood function (4.2) over the range $\mathcal{N} = \{x, x+1, x+2, \ldots\}$. To find this value, one must firstly maximise (4.2) for $p_j,\ j = 1, \ldots, t$, which, for a particular j, gives

$$\begin{aligned} \frac{n_j}{\hat{p}_j} &= \frac{N + R - n_j}{1 - \hat{p}_j} \\ \Rightarrow \hat{p}_j &= \frac{n_j}{N + R}. \end{aligned} \qquad (4.11)$$

Inserting (4.11) into (4.2) then gives the log profile likelihood. Maximising this for N over the range of $\mathcal{N}$ gives

$$\ln\{L(\hat{N}_t, \hat{p}_1(\hat{N}_t), \ldots, \hat{p}_t(\hat{N}_t)|(X, \vec{N}))\} = \max_{N\epsilon\mathcal{N}}\Bigg[\ln\left(\frac{N!}{(N-x)!}\right) + \sum_{j=1}^{t} n_j \ln(n_j) + \sum_{j=1}^{t}(N+R-n_j)\ln(N+R-n_j) - t(N+R)\ln(N+R)\Bigg]. \quad (4.12)$$

This is calculated iteratively and the value of N corresponding to the maximum value in this sequence thus becomes $\hat{N}_{M_{tp}}$, the M_{tp} MLE. The search is made feasible by assuming that the profile likelihood function is unimodal. Pickands & Raghavachari (1987) proved that the profile likelihood is unimodal under M_f and Goudie & Gormley (in submission) (see Appendix A) have proven unimodality under model M_{fp}. Goudie et al. (2007) have shown that the sub-model, model M_p, is unimodal for $R \geq 0$. Thus, the assumption is made that (4.12) is unimodal.
For a finite R, the variance estimator given by Darroch (1958, p. 352) is also suitable asymptotically under model M_{tp}. Thus, the variance estimator for $\hat{N}_{M_{tp}}$ used in §4.5 is

$$\text{var}\left(\hat{N}_{M_{tp}}\right) = \hat{N}_{M_{tp}}\left[\prod_{j=1}^{t}\frac{1}{(1-\hat{p_j})} + t - 1 - \sum_{j=1}^{t}\frac{1}{(1-\hat{p_j})}\right]^{-1}. \quad (4.13)$$

It is of interest here to see how this estimate of variance relates to the sample variance in the simulation work given in the next section.

4.4.2 Generalised Pathak Estimator

The generalised Pathak estimator is used under models M_t and M_{tp} here to test whether it performs well outside the confines of its original derivation. Recall from Chapter 2 the generalised Pathak estimator

$$\tilde{N}(x, \mathbf{n}, R) = x + \frac{a(x-1, \mathbf{n}, R)}{a(x, \mathbf{n}, R)}, \quad (4.14)$$

for $x = \max_j(n_j - R, 0), \ldots, \min(z, N)$, where $a(x, \mathbf{n}, R)$ is as given in (2.4).
As stated in Chapter 2, this follows from Berg's (1974) unbiased estimator for model M_f. It thus follows that this estimator is a conditionally unbiased estimator under model M_{tp}, conditioning on the sample sizes, $\mathbf{n}$. Under either M_{fp} or M_{tp} there is also the condition that $z \geq N$. This means that a unique, unbiased estimator of the variance of the Generalised Pathak estimator follows naturally from Berg (1974, eq^n 2.15) when the latter condition holds, as x is an observation from

a factorial series distribution. We will momentarily use the shorthand notation $\tilde{N}_x$ to represent (4.14), and $\tilde{N}_{x-1}$ to represent the corresponding estimator with $x-1$ distinct captures. The variance estimate, if $(x+R) > \max_j\{n_j\}$, given by Berg (1974), is

$$\hat{\text{var}}\left(\tilde{N}_x\right) = \left(\tilde{N}_x - x\right)\left(\tilde{N}_x - \tilde{N}_{x-1}\right). \tag{4.15}$$

This variance estimator, (4.15), has a problem case, when $(x+R) = \max_j\{n_j\}$, in that the term $\hat{N}_{x-1}$ contains a ratio with $a(x-1, \mathbf{n}, R)$ in the denominator, which is defined as being 0 in this situation. In this situation, we must use the final, non-numbered, equation given by Berg (1974) in the proof of his equation (2.15). This, he gives as

$$\begin{aligned} \hat{\text{var}}(\hat{N}_x) &= h_1(x)^2 + h_1(x) - h_1(x)h_1(x-1) \\ &= h_1(x)^2 + h_1(x) - h_2(x), \end{aligned} \tag{4.16}$$

where

$$h_\nu(x) = \begin{cases} \dfrac{a(x-\nu, \mathbf{n}, R)}{a(x, \mathbf{n}, R)} & \text{for } x \geq \nu,\ \nu = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned} \hat{\text{var}}\left(\tilde{N}_x\right) &= (\tilde{N}_x - x)^2 + (\tilde{N}_x - x) \\ &= 0 \end{aligned} \tag{4.17}$$

if $(x+R) = \max_j\{n_j\}$, as $h_2(x)$ equals zero when this condition holds.

4.4.3 Maximum Likelihood Estimator for M_p

As with the model M_{tp} MLE (p. 56), this is not a closed-form estimator, but instead seeks the maximum value of the likelihood function, $L(N)$, for $N \in \mathcal{N} = \{x, x+1, x+2, \ldots\}$, analogous to Goudie et al. (2007). It also requires the condition that z exceeds x in order for the likelihood estimate to remain finite (see Goudie et al. (2007, p. 245) for a proof). Thus, the log-likelihood of (N, p) is given by

$$\ell(N, p; z, x) = \ln\left(\frac{N!}{(N-x)!}\right) + z\ln(p) + (Nt + Rt - z)\ln(1-p).$$

As in Goudie et al. (2007) we find that $\ell(N, p; z, x)$ is maximised over $p \in [0, 1]$ when $\hat{p} = z/(Nt + Rt)$. Substituting this into the above log-likelihood function gives the log profile likelihood from which the M_p MLE for N is obtained. The maximum is obtained from iterative calculations through the range of $\mathcal{N}$ and the

M_p MLE, $\hat{N}_{M_p}$, is this maximum value of N. The lack of a closed-form estimator is unhelpful, but Goudie et al. (2007) showed that, under model M_p, the profile likelihood is unimodal and so iterations can stop once a turning point has been reached. They also gave a compact inequality for the value $k \in \mathcal{N}$ that gives the maximum value of the profile likelihood. The given inequality states that $\hat{N}_{M_p}$ "...is the smallest such k for which

$$\Delta h(kt + Rt) - \Delta h(kt + Rt - z) - \log\{(k - x + 1)/(k + 1)\}$$

is negative, where $h(k) = -k \log k$" and Δ denotes the forward finite difference operator. This goes some way to shortening the computational time involved.
An asymptotic variance estimate for the M_0 MLE is given by Darroch (1958) for model M_0, and Goudie et al. (2007) proved that it also applies to model M_p. Thus, the variance estimator for $\hat{N}_{M_p}$ is

$$\text{var}\left(\hat{N}_{M_p}\right) = \hat{N}_{M_p}\left[\frac{1}{(1-\hat{p})^t} + t - 1 - \frac{t}{(1-\hat{p})}\right]^{-1}. \tag{4.18}$$

4.4.4 M_p CUE

The M_p CUE was given in §2.3, namely

$$\tilde{N}_c = x + \frac{G(z, x-1, t, Rt)}{G(z, x, t, Rt)},$$

where $G(z, x, t, Rt)$ is a Gould-Hopper number, as given by (2.19) or in Gould & Hopper (1962). It is a function of the equal-catchability model sufficient statistics, namely z and x, but is used here under the more general case of time-heterogeneity. A unique, conditionally unbiased estimate of the variance follows naturally from the paper of Berg (1974) if $z \geq N$, as the conditional distribution of X given Z is an FSD. Firstly, as in §4.4.2, we let $\tilde{N}_{c-1}$ represent the corresponding CUE estimate if $(x - 1)$ distinct captures are observed. We also need a condition that $xt + Rt > z$ in order for the variance estimator to be finite. As in §4.4.2, we can define the variance estimator by

$$\text{var}\left(\tilde{N}_c\right) = \begin{cases} \left(\tilde{N}_c - x\right)\left(\tilde{N}_c - \tilde{N}_{c-1}\right) & x = \frac{z}{t} - R + 1, \ldots, z; \\ 0 & x = \frac{z}{t} - R, \end{cases} \tag{4.19}$$

by replacing the a-coefficients in §4.4.2 by the Gould-Hopper numbers and using the same methodology.

4.4.5 Petersen estimator

Finally, another estimator that was tried was a modification of the Petersen estimator. This estimator is sometimes referred to as the Lincoln-Petersen estimator, or the Lincoln Index. For a detailed history of the estimator, see Goudie & Goudie (2007).

Under model M_{tp} the ratio of x/N should be approximately equal to the corresponding ratio for the planted population, as there is no between animal heterogeneity. Thus, if we let n_R be the number of distinct R planted animals caught in the trial, then we get a basic estimator $\tilde{N}_p$, given by

$$\frac{x}{\tilde{N}_P} = \frac{n_R}{R}. \tag{4.20}$$

Rearranging gives an expression for $\tilde{N}_P$.

If the number of plants is relatively small with respect to N, the number of distinct planted animals caught, n_R, will also be relatively small with respect to N. Thus, there will be a non-negligible probability of having n_R=0, which leads to an infinite point population estimate. Thus, a modification, analogous to Bailey's (1951) binomial model modification, is made, giving a bias-corrected estimator

$$\hat{N}_P = \frac{(R+1)x}{n_R+1}. \tag{4.21}$$

To get an estimate of the variance, we again refer to Bailey (1951) and get an almost unbiased estimate, given by

$$\text{var}(\hat{N}_P) = \frac{x^2(R+1)(R-n_R)}{(n_R+1)^2(n_R+2)}. \tag{4.22}$$

The version used for mark-recapture, where $R = 0$, is a modification of this, using the animals captured in the first sample as the planted population, say $\tilde{R}$. It then calculates the number of these that were caught in the remaining samples, $n_{\tilde{R}}$. This is then compared with the number of distinct animals caught from the "target" population, n_x, this being those distinct animals that were not captured in the first sample, but subsequently caught. This gives an estimator

$$\hat{N}_P = \frac{(\tilde{R}+1)n_x}{n_{\tilde{R}}+1} + \tilde{R}, \tag{4.23}$$

where the addition of $\tilde{R}$ at the end is required as these animals are still part of the target population, despite the analysis being carried out as if they were planted.

4.5 M_{tp} Results

4.5.1 Method

Otis et al. (1978, Table N.2.b.) reported simulation results for the M_t MLE for a variety of fixed capture probabilities given in their Table N.2.a. Their intention was to use capture probabilities that cover a wide range of scenarios that the practitioner will encounter.

This approach is extended here to make the choices of capture probability more realistic, by assigning them a distribution. Following Huggins (2002) and Dorazio & Royle (2003) we assume that the capture probabilities are random variables from a beta distribution with fixed parameters α and β, which lead to different means and standard deviations. For the choice of means, further work not reported here suggested that for a capture probability mean of above $\mu = 0.2$, the difference in population estimation of the simulation trials between the different estimators is very marginal. Thus, α and β were chosen to reflect this. Further to this, the desired means were 0.05, 0.1 & 0.2, following Goudie & Ashbridge (2005, p. 1549) and Ashbridge & Goudie (2009, p. 7). For capture probabilities lower than 0.05, one is referred to the later chapter on sparse data, Chapter 5. An example of the capture probabilities produced under each distribution is contained in each of the results tables, Tables 4.2 – 4.7.

Values of α and β were also chosen to represent different levels of variability from the mean, and so were chosen to give standard deviations of 0.0125, 0.025, 0.05. Thus, nine different combinations of mean and standard deviation were used, and are given in Figure 4.1 and Table 4.1, which offer a wide range of possible scenarios that a practitioner would find out in the field. This should then allow for determination of the optimal estimator for each situation, where optimality is defined in terms of means and standard deviations.

In the simulations, one set of capture probabilities was drawn from the chosen distribution, and from this, 1000 realisations were simulated. For each realisation, a population estimate, a standard deviation estimate and a 95% confidence interval (based on a Normal approximation, which holds asymptotically (c.f. Fewster & Jupp (2009))) were calculated for each of the different estimators. These were then averaged, and it is these averages that are given in Tables 4.2 – 4.7. Also given is the proportion of the 1000 intervals that contain the true population size, N. Since a 95% confidence interval is calculated each time, the proportion of intervals containing N should be around 0.95. The final column gives the average width of the confidence interval throughout the 1000 realisations. The intervals are symmetrical under the Normal assumption, therefore are not truncated at the lower end.

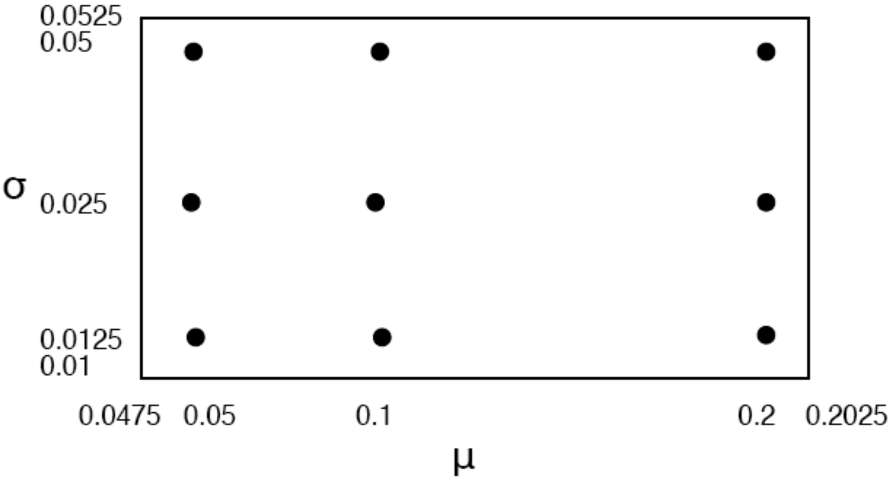


Figure 4.1: Diagram of the grid used for the Beta-generated probabilities.

μ	σ	(α, β)
0.05	0.0125	(15.15, 287.85)
	0.025	(3.75, 71.25)
	0.05	(0.9, 17.1)
0.1	0.0125	(57.5, 517.5)
	0.025	(14.3, 128.7)
	0.05	(3.5, 31.5)
0.2	0.0125	(204.6, 818.4)
	0.025	(51, 204)
	0.05	(12.6, 50.4)

Table 4.1: A reference table for the parameters of the beta-distributed capture probabilities used in the simulations

4.5.2 Results

The results, given in Tables 4.2 – 4.7, will be analysed from several perspectives, before general conclusions are given in §4.8. The first analysis will be to analyse the effect that total population size has on the quality of estimation. Secondly, analyses of the estimator results when 10 samples are carried out and when 5 samples are taken are given, and the relevant conclusions drawn. The optimal situation for the practitioner would be to plant as few additional animals as possible into the target population, and to carry out as few samples as possible, whilst still having a resultant estimate that can be assumed to have a small bias and a small standard deviation. This would result in the most cost and time effective trial. One requirement of the planted animals is that they do not alter the behaviour or capture probabilities of the target population. The addition of only a few planted animals would allow the practitioner to have more confidence in this assumption. Thus, it is with these goals that the decision was made to add 10 plants in each simulation, regardless of population size. This is to reflect the fact that the practitioner would not have knowledge of the true population, and so relating the optimal number of plants to the population would be impractical. Also, the plots shown in §2.5.2 are consistent with simulation work carried out under model M_{tp}. These show that the relative improvement in population estimation of each additional 5 plants becomes smaller.

We will further subdivide the results by mean capture probability with the aim of establishing the optimal estimator for any situation. This is a subdivision that is not possible to make in practice, but the hope is to find consistency across all sampling distributions or, failing that, to establish a pattern throughout the distributions sampled.

The N=50, t=5, μ=0.05 case

We see from Table 4.2 that the M_p CUE performs well under this model M_{tp} scenario, despite the estimator being a function of just X and Z. The mean point estimate from the M_p CUE is very close to the true population size in this case, with a sample standard deviation that is just over half the size of the point estimate. The estimated standard deviation underestimates the sample standard deviation by around 10% of N. This may explain why the proportion of 95% confidence intervals containing the true value is below 0.95, the target value. This suggests that the interval width is slightly smaller than it should be. This would lead to the practitioner having a falsely high level of confidence about the estimates, a situation one wishes to avoid.

In the first case the generalised Pathak mean estimate is much lower than N, mak-

(μ, σ)	Example p_js	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage proportion	Average width
(0.05, 0.0125)	0.0508	M_p CUE	50.26	28.77	24.03	0.817	94.19
	0.0652	M_p MLE	61.84	41.27	67.09	0.961	263.00
	0.0228	G. Pathak	36.91	77.08	28.42	0.698	110.72
	0.0432	M_{tp} MLE	56.99	37.81	64.53	0.946	252.95
	0.0679	Petersen	44.31	25.64	20.56	0.713	80.61
(0.05, 0.025)	0.0720	M_p CUE	49.04	29.52	23.35	0.796	91.53
	0.0508	M_p MLE	61.27	42.73	69.13	0.947	270.99
	0.0652	G. Pathak	48.60	22.94	17.86	0.849	70.02
	0.0228	M_{tp} MLE	57.45	38.23	66.55	0.962	260.87
	0.0432	Petersen	44.13	27.61	20.56	0.700	80.59
(0.05, 0.05)	0.0228	M_p CUE	48.28	28.28	22.40	0.816	87.81
	0.0058	M_p MLE	64.79	40.61	72.09	0.968	282.60
	0.0244	G. Pathak	50.25	18.79	15.88	0.883	62.27
	0.0150	M_{tp} MLE	56.08	31.62	38.84	0.954	152.25
	0.1327	Petersen	43.35	26.21	19.48	0.701	76.35
(0.1, 0.0125)	0.0972	M_p CUE	50.33	17.26	14.50	0.880	56.85
	0.1244	M_p MLE	53.35	20.85	23.40	0.951	91.73
	0.1232	G. Pathak	49.97	16.44	14.93	0.889	58.54
	0.0954	M_{tp} MLE	51.43	20.05	22.25	0.944	87.21
	0.1221	Petersen	49.64	23.15	17.03	0.778	66.76
(0.1, 0.025)	0.0734	M_p CUE	50.11	16.15	14.40	0.880	56.45
	0.0720	M_p MLE	53.56	21.58	23.27	0.950	91.22
	0.1256	G. Pathak	50.21	15.71	13.91	0.893	54.54
	0.1208	M_{tp} MLE	51.91	18.43	22.66	0.950	88.81
	0.1420	Petersen	49.87	24.06	17.32	0.788	67.90
(0.1, 0.05)	0.0734	M_p CUE	51.17	18.96	15.56	0.889	61.01
	0.2194	M_p MLE	54.13	20.04	24.95	0.956	97.79
	0.0720	G. Pathak	49.72	15.25	14.24	0.888	55.79
	0.1256	M_{tp} MLE	53.28	22.96	27.13	0.928	106.34
	0.0584	Petersen	48.96	23.46	16.76	0.768	65.71
(0.2, 0.0125)	0.2196	M_p CUE	49.96	7.36	6.77	0.919	26.54
	0.2017	M_p MLE	50.06	7.52	8.14	0.935	31.89
	0.1882	G. Pathak	50.00	7.16	6.62	0.933	25.98
	0.1896	M_{tp} MLE	49.62	7.51	8.03	0.924	31.50
	0.1861	Petersen	50.14	13.05	9.99	0.812	39.18
(0.2, 0.025)	0.2196	M_p CUE	49.76	7.04	6.84	0.924	26.81
	0.2017	M_p MLE	49.89	7.77	8.17	0.943	32.04
	0.2132	G. Pathak	50.06	6.94	6.55	0.944	25.68
	0.1882	M_{tp} MLE	49.77	7.10	8.00	0.928	31.37
	0.1711	Petersen	49.41	11.69	9.69	0.813	38.00
(0.2, 0.05)	0.2196	M_p CUE	50.44	7.90	7.14	0.917	28.00
	0.1420	M_p MLE	50.29	7.52	8.43	0.954	33.04
	0.1622	G. Pathak	50.05	6.42	6.39	0.937	25.03
	0.2017	M_{tp} MLE	49.82	7.36	8.23	0.946	32.26
	0.1395	Petersen	50.15	12.54	9.97	0.811	39.08

Table 4.2: Simulated results for 1000 realisations from a population $N = 50$, $R = 10$ and $t = 5$ with Beta-distributed capture probabilities.

(μ,σ)	Example p_js	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage proportion	Average width
(0.05, 0.0125)	0.0508	M_p CUE	49.88	14.63	14.14	0.901	55.42
	0.0652	M_p MLE	54.05	23.49	23.75	0.930	93.11
	0.0228	G. Pathak	50.67	18.85	16.04	0.886	62.86
	0.0432	M_{tp} MLE	53.35	19.47	22.59	0.939	88.56
	0.0679	Petersen	49.24	23.02	16.91	0.778	66.30
(0.05, 0.025)	0.0720	M_p CUE	50.41	16.76	14.84	0.899	58.17
	0.0508	M_p MLE	54.33	24.24	24.28	0.956	95.19
	0.0652	G. Pathak	49.38	15.87	14.07	0.878	55.14
	0.0228	M_{tp} MLE	53.35	23.17	24.75	0.935	97.03
	0.0432	Petersen	49.46	22.80	17.13	0.795	67.15
(0.05, 0.05)	0.0228	M_p CUE	51.51	16.78	14.63	0.912	57.34
	0.0058	M_p MLE	56.59	23.70	25.83	0.967	101.27
	0.0244	G. Pathak	50.11	15.15	14.21	0.903	55.67
	0.0150	M_{tp} MLE	51.65	18.89	22.42	0.928	87.87
	0.1327	Petersen	49.49	25.03	16.38	0.782	64.21
(0.1, 0.0125)	0.0972	M_p CUE	50.04	7.52	7.19	0.922	28.19
	0.1244	M_p MLE	50.21	7.90	8.56	0.944	33.56
	0.1232	G. Pathak	49.99	7.67	7.22	0.928	28.28
	0.0954	M_{tp} MLE	49.62	7.46	8.24	0.938	32.30
	0.1221	Petersen	50.40	14.84	10.68	0.828	41.86
(0.1, 0.025)	0.0734	M_p CUE	49.88	7.42	7.11	0.925	27.89
	0.0720	M_p MLE	50.28	7.76	8.54	0.945	33.50
	0.1256	G. Pathak	49.95	7.22	6.90	0.933	27.04
	0.1208	M_{tp} MLE	49.64	7.75	8.35	0.937	32.74
	0.1420	Petersen	49.52	12.82	10.23	0.808	40.10
(0.1, 0.05)	0.0734	M_p CUE	50.48	7.77	7.32	0.928	28.70
	0.2194	M_p MLE	50.81	8.27	9.03	0.952	35.40
	0.0720	G. Pathak	49.94	9.93	9.26	0.924	36.33
	0.1256	M_{tp} MLE	49.33	7.94	8.50	0.923	33.30
	0.0584	Petersen	50.03	12.77	10.42	0.835	40.84
(0.2, 0.0125)	0.2196	M_p CUE	50.04	2.86	2.83	0.946	11.08
	0.2017	M_p MLE	49.52	2.88	2.91	0.930	11.40
	0.1882	G. Pathak	50.05	2.41	2.68	0.959	10.54
	0.1896	M_{tp} MLE	49.34	2.83	2.86	0.921	11.21
	0.1861	Petersen	50.05	5.99	4.31	0.704	16.91
(0.2, 0.025)	0.2196	M_p CUE	50.01	2.89	2.79	0.938	10.95
	0.2017	M_p MLE	49.61	2.91	2.94	0.933	11.52
	0.2132	G. Pathak	50.04	3.03	2.90	0.955	11.41
	0.1882	M_{tp} MLE	49.39	2.87	2.89	0.924	11.33
	0.1711	Petersen	50.14	5.97	4.24	0.673	16.63
(0.2, 0.05)	0.2196	M_p CUE	50.32	2.91	2.88	0.948	11.31
	0.1420	M_p MLE	49.63	2.99	2.98	0.944	11.68
	0.1622	G. Pathak	49.99	3.11	3.08	0.960	12.13
	0.2017	M_{tp} MLE	49.32	2.93	2.89	0.911	11.32
	0.1395	Petersen	49.98	6.13	4.17	0.666	16.34

Table 4.3: Simulated results for 1000 realisations from a population $N = 50$, $R = 10$ and $t = 10$ with Beta-distributed capture probabilities.

ing it the most biased of all the estimators in this trial. However, the mean estimates improve as σ increases, to the point where it is the least biased estimator in the third trial. In the first trial the estimated standard deviation of the point estimate drastically underestimates the sample standard deviation. For the other two cases the estimated standard deviation is close to, but an underestimate of, its sample equivalent. This should explain why the coverage interval is much lower than 0.95 in the first case and a better approximation to 0.95 in the other two cases.

The Petersen-type estimator has a mean point estimate that is below N by over 10% and a sample standard deviation that is just above 50% of the mean point population estimate in all three trials. The mean estimated standard deviation underestimates the sample standard deviation by about a fifth each time, which results in the coverage proportion being less than 75%, which suggests that the average width of the confidence interval is too small in a lot of simulations. This suggests that the standard deviation estimator needs to be improved in this case.

Both MLEs have the opposite problem to the other 3 estimators in that their mean point estimates overestimate the true population and their standard deviation estimates are larger than their sample equivalents. With the exception of the third M_{tp} MLE trial, these estimated standard deviations would result in a negative lower bound for their average coverage values much higher than the non-MLE estimators. This would imply that their lower bound in these cases would be the uninformative value, x, with an upper bound that is more than double the true population value, N. As a result of the higher standard deviations, the MLEs have coverage values much larger than the non-MLE estimators. Practitioners, if animal conservation is foremost in their minds, would often favour an underestimate of true population, as this would not lead them into a false belief of abundance.

As the means of both MLEs overestimate N, they cannot be recommended for use. The M_p CUE, the generalised Pathak estimator and the Petersen-type estimator have a tendency to underestimate N on average, but the favoured estimator is the M_p CUE, as it is the most consistent estimator, with its mean point estimate in the sparsest data case (μ=0.05, σ=0.0125) being within 2% of N. It also has a much lower sample standard deviation in this case than the generalised Pathak estimator, and a more reliable estimated standard deviation. The reason for its good performance in a model with time-heterogeneity is probably attributable to the low capture probabilities. Coupled with the small population, this will result in few captures in each sample. Thus, there will be little variation between the samples, resulting in the homogeneous estimator's good performance.

The N=50, t=5, μ=0.1 case

The generalised Pathak has a mean point estimate that is within ± 1 of the true population, N, in all three trials, whilst the M_p CUE and Petersen-type mean estimates deviate by more than ±1 in only the $\sigma = 0.05$ trial. The confidence intervals for the 1000 realisations always contain N less than 90% of the time, however, which is lower than the target level. This is a result of the standard deviation estimators underestimating the sample standard deviation of the point estimates.

The mean of the M_{tp} MLE here is seen to overestimate the population size in all three trials, with no improvement as σ increases. Its estimated standard deviation also overestimates the sample standard deviation in all three trials, yet its coverage proportion never exceeds 0.95.

The M_p MLE overestimates the true population in all three trials, but is an improvement on the $\mu = 0.05$ case. The mean estimated standard deviation still overestimates the sample standard deviation, which implies that the confidence interval associated with the mean M_p MLE estimate is too conservative. The coverage proportion is above 0.95 for each trial, which suggests that there is room for improvement in the standard deviation estimator.

Thus, overall, the generalised Pathak estimator is preferable in this case.

The N=50, t=5, μ=0.2 case

All estimators have mean point estimates within ±0.5 of the true population size in this case, with the exception of one Petersen-type estimate. The mean estimated standard deviations are lower than the corresponding estimate when $\mu = 0.1$, sometimes more than a third smaller. This suggests that if the capture probability mean is 0.2, the precision of the estimates from any estimator is very high, and there is little to distinguish between them. The MLEs' mean estimated standard deviations again slightly overestimate their sample standard deviations, but the coverage proportions for these estimators are still generally below 0.95. The estimated and sample standard deviations of the Petersen-type estimator are again larger than the rest, but it does not have a coverage proportion to justify this increase. Thus, the Petersen estimator again has a good mean point estimate but is poor in terms of interval estimation. There is little to distinguish between the rest of the estimators, however, in terms of mean point estimate or interval estimate.

However, on the basis of having the mean point estimate closest to N in each trial, and having the lowest estimated and sample standard deviations, the recommended estimator is the generalised Pathak estimator.

The N=50, t=10, μ=0.05 case

The M_p CUE here has the least biased mean point estimate of all the estimators for $\sigma = 0.0125$, but it increases as σ increases to the point where it overestimates by roughly 3% when $\sigma = 0.05$. A larger capture probability standard deviation means that there is liable to be a stronger time-heterogeneity element between the samples. The sample standard deviation is lower than in the $t = 5$ case and the mean estimated standard deviation is much closer to the sample standard deviation when $t = 10$. This leads to improved coverage proportions for each trial as the variability in the data is being captured better by the standard deviation estimator. Hence, the interval estimates are wide enough to contain N on more occasions. The coverage proportion is around 90%, which is still somewhat below the desired level of 95%.

The generalised Pathak estimator here estimates N with almost no bias, but still has a coverage proportion that only once manages to exceed 90%. This is because the mean estimated standard deviation is underestimating the sample standard deviation. The difference between the two is very small, however, and so is satisfactory.

The Petersen-type estimator has a mean population estimate that is within unity of N, but it too suffers from having a standard deviation estimate that is too low, making the confidence intervals too narrow to contain N the desired 95% of the time.

The mean of the M_{tp} MLE when $t = 10$ is closer to N than when $t = 5$, but still fails to get within 3% of N. It is evident here that the M_{tp} MLE is better than the M_p MLE in terms of mean point estimate and also for sample standard deviation. The M_p MLE has, in one trial, a lower estimated standard deviation than the M_{tp} MLE, but overall the M_{tp} MLE is the superior of the two MLEs.

Considering all this, however, there is no clear recommendation. The generalised Pathak estimator and M_p CUE both perform well when considering the mean point population estimate and estimated standard deviation. The one flaw with the CUE is that it may be becoming more biased as the heterogeneity among the animals is increasing.

The N=50, t=10, μ=0.1 case

In this situation, there is very little to choose between the estimators, as all estimators have mean population estimates within ± 1 of the true population size of 50. The Petersen-type estimator has a coverage proportion that is too low to be considered acceptable, despite having the highest mean estimated standard deviation. The other estimators all have coverages of over 90%. The MLEs' mean estimated standard deviations overestimate their sample standard deviations consistently. For the other estimators, their mean estimated standard deviation is always below their sample standard deviation counterpart. However, by virtue of its very marginal overall optimality in terms of mean point estimate, the generalised Pathak estimator is proposed for this case.

The N=50, t=10, μ=0.2 case

All the estimators except from the M_{tp} MLE have average estimates that equal the true population within rounding, and all estimators have a very low estimated standard deviation.

As has been noticed in some of the other cases above, the M_p CUE has a mean point estimate very close to N for σ=0.0125 , but has a mean estimate that overestimates when $\sigma = 0.05$, when the time-heterogeneity becomes more pronounced. In this last trial, it still gives an estimate with an accuracy within rounding of N.

The generalised Pathak estimator gives mean point estimates that are very accurate in all three trials. In one trial it has a mean estimated standard deviation above its corresponding sample standard deviation.

The M_{tp} MLE underestimates N in every trial, as in the μ=0.1 case above it. The mean population size estimate by the M_p MLE is now lower than N, although it rounds up to the true value.

The Petersen-type estimator's average standard deviation is larger than the rest, but is still too low for its interval estimate to contain N in 95% of cases.

As the generalised Pathak estimator's mean estimate is the most consistent out of all the estimators in this case, this is the proposed estimator.

When N=100

For a small, fixed mean capture probability, $\mu = 0.05$, the M_p CUE appears to lose precision as the capture probability standard deviation increases, making the time-heterogeneity more pronounced. This is similar to what was observed in Tables 4.2–4.3 and is understandable, as the estimator is designed to be unbiased under homogeneous capturing. A high capture probability standard deviation causes an overestimate of the population size. In the simulations, column 2 of Tables 4.2 – 4.7 give sample capture probabilities randomly generated from the associated Be(α, β) distributions. For the $(\mu, \sigma) = (0.05, 0.05)$ trial with $t = 5$, for example, one sample had a capture probability of just 0.1%, whilst another had a capture probability of 7%. This represents a strong time-heterogeneity effect between samples, giving a major departure from the assumption of a constant capture probability, on which the M_p CUE and the M_p MLE are based. This results in a deterioration of precision in some cases. The effect of this violation appears to be reduced by increasing the number of samples, resulting in more data being available.

The generalised Pathak estimator has a mean estimate, in all the cases shown, within unity to the true population size, which is very desirable. It is only when $\mu = 0.05$ that there is some deviation away from the true population size. Its mean estimated standard deviation also closely estimates the sample standard deviation in most trials, especially when μ and/or t increase. The average coverage proportion of the estimator is generally below the desired 0.95 level when $\mu = 0.05$ but improves with μ and t. Thus, the standard deviation estimator does well in most

(μ, σ)	Example p_js	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage proportion	Average width
(0.05, 0.0125)	0.0508	M_p CUE	98.77	51.38	40.55	0.825	158.97
	0.0652	M_p MLE	121.73	76.16	92.98	0.950	364.49
	0.0228	G. Pathak	99.26	47.45	38.34	0.852	150.35
	0.0432	M_{tp} MLE	117.47	72.64	89.55	0.941	351.05
	0.0679	Petersen	93.06	54.04	45.03	0.734	176.52
(0.05, 0.025)	0.0720	M_p CUE	102.91	58.33	44.13	0.851	172.97
	0.0508	M_p MLE	123.31	79.28	100.50	0.952	393.97
	0.0652	G. Pathak	99.76	46.98	39.75	0.851	155.82
	0.0228	M_{tp} MLE	119.66	80.22	99.48	0.942	389.95
	0.0432	Petersen	89.85	50.15	42.43	0.719	166.34
(0.05, 0.05)	0.0228	M_p CUE	106.38	54.29	44.02	0.891	172.55
	0.0058	M_p MLE	125.97	75.58	102.46	0.956	401.63
	0.0244	G. Pathak	100.19	48.52	39.25	0.847	153.86
	0.0150	M_{tp} MLE	114.90	68.28	98.74	0.949	387.08
	0.1327	Petersen	87.29	50.66	39.63	0.717	155.36
(0.1, 0.0125)	0.0972	M_p CUE	100.08	23.53	22.51	0.892	88.23
	0.1244	M_p MLE	105.89	29.57	30.87	0.949	121.03
	0.1232	G. Pathak	99.59	23.15	22.51	0.907	88.23
	0.0954	M_{tp} MLE	103.80	27.52	30.74	0.948	120.50
	0.1221	Petersen	99.13	40.50	33.59	0.813	131.66
(0.1, 0.025)	0.0734	M_p CUE	100.30	24.20	23.00	0.901	90.14
	0.0720	M_p MLE	104.84	26.94	31.13	0.942	122.01
	0.1256	G. Pathak	99.25	24.64	21.51	0.895	84.30
	0.1208	M_{tp} MLE	104.66	29.80	31.20	0.939	122.30
	0.1420	Petersen	99.24	45.11	33.89	0.805	132.86
(0.1, 0.05)	0.0734	M_p CUE	103.57	30.53	25.06	0.929	98.22
	0.2194	M_p MLE	106.24	31.03	33.55	0.956	131.50
	0.0720	G. Pathak	100.37	17.92	16.76	0.920	65.68
	0.1256	M_{tp} MLE	105.81	32.92	34.61	0.941	135.68
	0.0584	Petersen	100.25	50.87	34.80	0.801	136.42
(0.2, 0.0125)	0.2196	M_p CUE	99.74	11.04	10.30	0.923	40.38
	0.2017	M_p MLE	100.27	10.40	11.50	0.947	45.10
	0.1882	G. Pathak	99.73	10.85	10.54	0.941	41.31
	0.1896	M_{tp} MLE	99.60	10.86	11.22	0.946	43.98
	0.1861	Petersen	100.25	24.54	20.10	0.836	78.78
(0.2, 0.025)	0.2196	M_p CUE	99.90	10.79	10.37	0.930	40.67
	0.2017	M_p MLE	100.10	11.01	11.40	0.940	44.70
	0.2132	G. Pathak	100.04	10.0	10.18	0.950	39.92
	0.1882	M_{tp} MLE	100.15	10.54	11.40	0.962	44.70
	0.1711	Petersen	98.47	22.38	19.08	0.825	74.79
(0.2, 0.05)	0.2196	M_p CUE	101.62	11.10	10.85	0.953	42.53
	0.1420	M_p MLE	101.40	11.29	11.87	0.963	46.55
	0.1622	G. Pathak	99.57	11.63	11.26	0.930	44.10
	0.2017	M_{tp} MLE	99.51	10.63	11.40	0.942	44.70
	0.1395	Petersen	100.59	23.27	20.25	0.863	79.37

Table 4.4: Simulated results for 1000 realisations from a population $N = 100$, $R = 10$ and $t = 5$ with Beta-distributed capture probabilities.

(μ, σ)	Example p_js	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage proportion	Average width
(0.05, 0.0125)	0.0508	M_p CUE	100.24	23.64	21. 53	0.908	84.40
	0.0652	M_p MLE	103.70	26.15	27.66	0.957	108.41
	0.0228	G. Pathak	99.40	27.60	25.56	0.886	100.20
	0.0432	M_{tp} MLE	102.84	24.37	27.26	0.940	106.85
	0.0679	Petersen	100.23	45.43	33.96	0.808	133.12
(0.05, 0.025)	0.0720	M_p CUE	100.77	31.11	27.91	0.895	109.40
	0.0508	M_p MLE	105.99	32.45	32.94	0.946	129.14
	0.0652	G. Pathak	99.83	20.82	20.11	0.920	78.85
	0.0228	M_{tp} MLE	105.35	32.18	32.18	0.942	123.48
	0.0432	Petersen	103.21	44.55	30.49	0.847	119.50
(0.05, 0.05)	0.0228	M_p CUE	104.82	34.03	29.77	0.904	116.69
	0.0058	M_p MLE	117.01	39.54	44.03	0.976	172.60
	0.0244	G. Pathak	99.49	12.46	12.71	0.932	49.87
	0.0150	M_{tp} MLE	102.81	24.25	27.40	0.945	107.39
	0.1327	Petersen	100.09	44.18	34.49	0.818	135.20
(0.1, 0.0125)	0.0972	M_p CUE	99.64	11.75	11.46	0.936	44.93
	0.1244	M_p MLE	101.07	10.55	11.31	0.955	44.35
	0.1232	G. Pathak	100.18	11.39	10.85	0.942	42.56
	0.0954	M_{tp} MLE	99.69	10.79	11.14	0.948	43.68
	0.1221	Petersen	99.73	25.12	20.79	0.833	81.50
(0.1, 0.025)	0.0734	M_p CUE	100.22	10.08	9.96	0.934	39.04
	0.0720	M_p MLE	100.36	10.44	11.18	0.958	43.84
	0.1256	G. Pathak	100.52	10.48	10.38	0.954	40.69
	0.1208	M_{tp} MLE	100.45	10.43	11.05	0.951	43.31
	0.1420	Petersen	100.93	23.98	19.89	0.835	77.96
(0.1, 0.05)	0.0734	M_p CUE	101.74	14.24	13.92	0.941	54.57
	0.2194	M_p MLE	102.43	13.63	14.49	0.965	56.81
	0.0720	G. Pathak	99.61	11.74	10.92	0.936	42.80
	0.1256	M_{tp} MLE	99.49	9.35	9.96	0.950	39.05
	0.0584	Petersen	99.81	20.67	17.64	0.843	69.15
(0.2, 0.0125)	0.2196	M_p CUE	100.11	3.72	3.70	0.940	14.50
	0.2017	M_p MLE	99.47	3.99	4.06	0.943	15.93
	0.1882	G. Pathak	100.23	4.04	4.01	0.961	15.70
	0.1896	M_{tp} MLE	99.43	4.31	4.27	0.938	16.74
	0.1861	Petersen	99.80	10.53	8.27	0.687	32.41
(0.2, 0.025)	0.2196	M_p CUE	99.95	3.74	3.75	0.947	14.70
	0.2017	M_p MLE	99.58	4.50	4.64	0.954	18.20
	0.2132	G. Pathak	99.91	3.67	3.75	0.967	14.72
	0.1882	M_{tp} MLE	99.32	4.25	4.17	0.924	16.33
	0.1711	Petersen	99.67	11.54	9.28	0.732	36.38
(0.2, 0.05)	0.2196	M_p CUE	100.35	4.65	4.48	0.937	17.58
	0.1420	M_p MLE	100.09	4.01	4.11	0.948	16.13
	0.1622	G. Pathak	100.04	4.56	4.32	0.956	16.93
	0.2017	M_{tp} MLE	99.68	4.13	4.35	0.959	17.04
	0.1395	Petersen	99.82	14.15	11.00	0.784	43.14

Table 4.5: Simulated results for 1000 realisations from a population $N = 100$, $R = 10$ and $t = 10$ with Beta-distributed capture probabilities.

cases, with an improvement perhaps possible when the mean capture probability is small.

The MLEs have a tendency to overestimate the true population size when the capture probabilities are small, especially when $t = 5$, and have very wide confidence intervals that may not offer the practitioner too much additional information. The wide interval can be narrowed by carrying out 10 samples rather than 5, but this may not be possible in practice. As with the cases when $N = 50$, the mean estimated standard deviations overestimate the sample standard deviations, but this overestimation decreases as μ and/or t increase. As the mean capture probability increases, the difference in the summary statistics of each of the estimators becomes small.

The Petersen-type estimator also performs well when the mean estimate is considered, but its average estimated and sample standard deviations are generally higher than the other estimators' standard deviations. Despite this, the coverage proportion for the Petersen-type estimator is consistently lower than that of the other estimators.

Thus, the proposed estimator is the generalised Pathak estimator, as this is the most consistently optimal estimator.

When N=250

The M_p CUE performs well in most of the trials when $N = 250$. When $t = 5,\ \mu = 0.1$ and $\sigma = 0.05$ the mean point estimate is 266, which stands out as being a large overestimate. In every other trial, the mean point estimate is within ± 10 of N. When $\mu = 0.05$ the mean point estimate increases in bias between $t = 5$ and $t = 10$, which is not evident in the previous tables. For μ=0.1 and μ=0.2 there is an improvement in mean point estimate from $t = 5$ and $t = 10$. The sample and mean estimated standard deviations decrease both as t is increased from 5 to 10 and as μ increases, which is consistent with the previous tables.

The generalised Pathak estimator is seen in every scenario to have a mean population estimate very close to the true population size and, in all scenarios except the $\mu = 0.05,\ t = 5$ cases and one $\mu = 0.05$ and $t = 10$ case, has a coverage proportion greater than 0.9. For the $\mu = 0.05,\ t = 5$ case, the mean estimated standard deviation underestimates the sample standard deviation by between 8 and 15%. Apart from that, despite underestimating by 9% on one other occasion, the mean estimated standard deviations appear to estimate the sample standard deviation well, and lead to coverage proportions of over 0.9.

The M_{tp} MLE performs poorly when $t = 5$ and $\mu = 0.05$ but has a mean point estimate that is within 1% in most other trials. The mean estimated standard deviation

(μ, σ)	Example	Estimator	Mean	Sample	Mean est.	Coverage	Average
	p_js		Estimate	std dev.	std dev.	proportion	width
(0.05, 0.0125)	0.0508	M_p CUE	248.26	76.59	69.14	0.896	271.02
	0.0652	M_p MLE	272.60	103.17	100.72	0.946	394.83
	0.0228	G. Pathak	249.84	91.84	76.03	0.869	298.06
	0.0432	M_{tp} MLE	278.38	138.39	122.94	0.945	481.91
	0.0679	Petersen	232.55	138.58	105.66	0.772	414.20
(0.05, 0.025)	0.0720	M_p CUE	255.97	89.23	78.07	0.893	306.04
	0.0508	M_p MLE	310.47	182.23	173.58	0.958	680.43
	0.0652	G. Pathak	245.90	76.34	70.26	0.890	275.44
	0.0228	M_{tp} MLE	280.42	144.56	139.42	0.939	546.54
	0.0432	Petersen	212.72	145.72	113.73	0.699	445.83
(0.05, 0.05)	0.0228	M_p CUE	254.87	78.44	73.15	0.915	286.76
	0.0058	M_p MLE	324.92	119.40	126.62	0.991	496.37
	0.0244	G. Pathak	247.63	68.18	62.62	0.890	245.49
	0.0150	M_{tp} MLE	288.99	147.00	159.95	0.946	627.01
	0.1327	Petersen	250.78	134.96	96.07	0.804	376.59
(0.1, 0.0125)	0.0972	M_p CUE	252.69	41.75	41.87	0.937	164.15
	0.1244	M_p MLE	258.93	48.27	47.62	0.955	186.66
	0.1232	G. Pathak	250.09	38.74	37.90	0.926	148.59
	0.0954	M_{tp} MLE	253.71	43.43	44.77	0.951	175.49
	0.1221	Petersen	241.71	102.23	80.17	0.809	314.25
(0.1, 0.025)	0.0734	M_p CUE	253.50	48.38	44.04	0.921	172.65
	0.0720	M_p MLE	260.33	54.21	55.06	0.953	215.84
	0.1256	G. Pathak	250.44	47.48	43.00	0.913	168.57
	0.1208	M_{tp} MLE	255.19	44.31	45.91	0.953	179.97
	0.1420	Petersen	244.76	103.57	84.32	0.825	330.55
(0.1, 0.05)	0.0734	M_p CUE	266.35	46.12	44.66	0.956	175.06
	0.2194	M_p MLE	255.22	31.20	33.73	0.967	132.24
	0.0720	G. Pathak	250.24	42.05	40.49	0.933	158.73
	0.1256	M_{tp} MLE	252.65	48.69	48.08	0.932	188.49
	0.0584	Petersen	254.97	106.32	75.40	0.848	295.57
(0.2, 0.0125)	0.2196	M_p CUE	250.64	18.33	18.10	0.948	70.95
	0.2017	M_p MLE	250.61	17.01	17.51	0.957	68.63
	0.1882	G. Pathak	249.07	16.89	17.09	0.943	67.00
	0.1896	M_{tp} MLE	249.41	17.61	17.71	0.942	69.42
	0.1861	Petersen	250.00	63.54	50.26	0.840	197.03
(0.2, 0.025)	0.2196	M_p CUE	251.08	15.61	16.12	0.959	63.17
	0.2017	M_p MLE	250.77	15.64	16.58	0.960	65.01
	0.2132	G. Pathak	249.42	18.50	18.18	0.938	71.25
	0.1882	M_{tp} MLE	249.40	17.35	17.63	0.948	69.12
	0.1711	Petersen	248.71	63.25	53.37	0.828	209.20
(0.2, 0.05)	0.2196	M_p CUE	253.13	16.34	15.81	0.943	61.99
	0.1420	M_p MLE	253.50	20.03	19.84	0.953	77.76
	0.1622	G. Pathak	250.26	18.65	17.99	0.932	70.50
	0.2017	M_{tp} MLE	250.55	18.53	18.26	0.952	71.57
	0.1395	Petersen	249.19	50.27	39.31	0.781	154.08

Table 4.6: Simulated results for 1000 realisations from a population $N = 250$, $R = 10$ and $t = 5$ with Beta-distributed capture probabilities.

(μ, σ)	Example p_js	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage proportion	Average width
(0.05, 0.0125)	0.0508	M_p CUE	252.80	38.89	37.41	0.935	146.64
	0.0652	M_p MLE	258.83	43.90	42.58	0.959	166.90
	0.0228	G. Pathak	251.30	38.52	37.89	0.938	148.53
	0.0432	M_{tp} MLE	249.94	39.61	40.26	0.932	157.80
	0.0679	Petersen	248.42	106.92	83.43	0.807	327.05
(0.05, 0.025)	0.0720	M_p CUE	257.66	38.69	37.03	0.956	145.16
	0.0508	M_p MLE	261.33	44.93	47.14	0.963	184.77
	0.0652	G. Pathak	248.93	45.86	41.49	0.917	162.66
	0.0228	M_{tp} MLE	258.15	64.92	68.04	0.948	266.70
	0.0432	Petersen	251.32	119.03	87.84	0.822	344.34
(0.05, 0.05)	0.0228	M_p CUE	259.37	37.09	37.51	0.964	147.02
	0.0058	M_p MLE	295.40	59.82	61.00	0.986	239.13
	0.0244	G. Pathak	247.01	63.01	59.95	0.898	235.00
	0.0150	M_{tp} MLE	250.18	21.95	21.62	0.938	84.74
	0.1327	Petersen	249.66	93.99	79.80	0.826	312.82
(0.1, 0.0125)	0.0972	M_p CUE	251.80	18.30	18.62	0.950	72.98
	0.1244	M_p MLE	249.94	16.45	16.76	0.946	65.69
	0.1232	G. Pathak	250.01	17.37	17.50	0.959	68.61
	0.0954	M_{tp} MLE	250.29	18.07	19.20	0.958	75.26
	0.1221	Petersen	246.06	57.61	51.47	0.835	201.76
(0.1, 0.025)	0.0734	M_p CUE	253.04	17.21	17.32	0.957	67.91
	0.0720	M_p MLE	251.35	20.63	21.29	0.958	83.47
	0.1256	G. Pathak	250.96	18.20	17.71	0.941	69.44
	0.1208	M_{tp} MLE	250.11	18.87	18.77	0.951	73.58
	0.1420	Petersen	249.38	61.90	49.74	0.836	194.98
(0.1, 0.05)	0.0734	M_p CUE	251.74	14.07	14.24	0.951	55.82
	0.2194	M_p MLE	258.35	18.33	19.67	0.972	77.10
	0.0720	G. Pathak	249.52	20.75	21.00	0.939	82.29
	0.1256	M_{tp} MLE	248.82	13.81	13.72	0.929	53.77
	0.0584	Petersen	249.75	61.65	53.16	0.835	208.37
(0.2, 0.0125)	0.2196	M_p CUE	250.16	5.80	6.09	0.963	23.89
	0.2017	M_p MLE	249.66	6.62	6.74	0.950	26.41
	0.1882	G. Pathak	250.05	5.96	6.18	0.953	24.19
	0.1896	M_{tp} MLE	249.10	6.60	6.78	0.941	26.56
	0.1861	Petersen	249.19	27.65	20.37	0.662	79.85
(0.2, 0.025)	0.2196	M_p CUE	250.40	6.15	6.23	0.952	24.41
	0.2017	M_p MLE	249.89	6.35	6.48	0.948	25.41
	0.2132	G. Pathak	250.27	6.70	6.84	0.956	26.80
	0.1882	M_{tp} MLE	249.27	6.17	6.19	0.955	24.25
	0.1711	Petersen	249.45	25.37	18.39	0.616	72.11
(0.2, 0.05)	0.2196	M_p CUE	250.51	7.86	7.69	0.939	30.13
	0.1420	M_p MLE	250.62	6.75	6.96	0.957	27.27
	0.1622	G. Pathak	250.33	9.03	9.19	0.955	36.02
	0.2017	M_{tp} MLE	249.80	7.65	7.56	0.946	29.63
	0.1395	Petersen	250.27	31.23	24.35	0.741	95.44

Table 4.7: Simulated results for 1000 realisations from a population $N = 250$, $R = 10$ and $t = 10$ with Beta-distributed capture probabilities.

is lower than the sample standard deviation on quite a few trials, and appears to estimate the standard deviation quite well. This is to be expected, as the variance estimator (4.13) holds asymptotically.

The M_p MLE has a poor mean estimate when $\mu = 0.05$, and most of the trials when $\mu = 0.1$. There is no clear relationship between its sample and mean estimated standard deviations. Based on its mean point population estimate, it cannot be recommended when $N = 250$.

The Petersen-type estimator has, in most scenarios, a mean population estimate very close to the true population size. However, its mean estimated and sample standard deviations are often much larger than the corresponding standard deviation estimates for the other estimators, but still lead to poor coverage proportions.

With this set of cases, the generalised Pathak estimator is a consistently good estimator in the cases where the other estimators perform poorly. Thus, again, the generalised Pathak estimator is proposed as being optimal under model M_{tp} when $N = 250$.

4.6 A brief analysis under model M_t

We will now have a brief look at how far these results remain valid in the non-plant case, model M_t. It is of interest to note whether the generalised Pathak estimator or the M_p CUE could be recommended for a different model to that for which they were originally designed, and whether either could be preferred to the most commonly used estimator, the M_t MLE. The analysis for this will again be done via simulation; firstly, using capture probabilities generated from the beta distributions given in Table 4.1, and then in a similar vein to Otis et al.'s (1978, p. 126) Table N.2.b. Thus, Table 4.8 below should be compared to Table 4.2 above. In this table, 1000 realisations from each distribution are simulated to provide the mean estimates. Otis et al.'s (1978) Table N.2.b. is recreated under model M_t in Table 4.9. Table 4.9 compares the values of the M_t MLE given in the Otis et al table (using resimulated values) with that of the generalised Pathak estimator, and the M_0 CUE and M_0 MLE. The number of realisations for each trial are the same as those specified by Otis et al. (1978) in their Table N.2.b., column 7. The summary statistics given in Table 4.9 are mean and average standard deviation estimates, as well as some interval summary statistics. Thus, columns 6, 7, 8 and 9 of Table 4.9 should be compared to columns 3, 4, 5 and 6 respectively of Otis et al.'s Table N.2.b. Column 10 of Table 4.9 can be compared to Otis et al.'s Table N.2.b by evaluating $(2 \times 1.96) \cdot \text{Ave}\sqrt{\hat{\text{Var}}(\hat{N})}$, where $\text{Ave}\sqrt{\hat{\text{Var}}(\hat{N})}$ is Otis et al.'s column 5.

(μ, σ)	Example p_js	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage proportion	Average width
(0.05, 0.0125)	0.0508	M_0 CUE	40.28	22.69	20.74	0.690	81.32
	0.0652	M_0 MLE	47.36	26.08	34.62	0.830	139.58
	0.0228	G. Pathak	40.21	22.53	20.81	0.653	79.68
	0.0432	M_t MLE	43.83	25.30	31.05	0.772	121.70
	0.0679	Petersen	33.39	16.80	17.03	0.530	66.75
(0.05, 0.025)	0.0720	M_0 CUE	44.37	25.80	22.82	0.736	89.44
	0.0508	M_0 MLE	53.41	29.66	37.98	0.867	152.69
	0.0652	G. Pathak	40.67	22.62	20.88	0.690	81.83
	0.0228	M_t MLE	38.21	22.53	28.27	0.720	110.81
	0.0432	Petersen	34.47	17.14	17.30	0.565	67.81
(0.05, 0.05)	0.0228	M_0 CUE	41.46	23.59	21.84	0.725	85.60
	0.0058	M_0 MLE	67.50	39.28	41.46	0.933	177.67
	0.0244	G. Pathak	49.10	23.48	17.61	0.810	69.03
	0.0150	M_t MLE	34.90	17.63	20.50	0.591	80.37
	0.1327	Petersen	38.63	19.32	18.54	0.631	72.67
(0.1, 0.0125)	0.0972	M_0 CUE	49.32	20.76	16.98	0.823	68.70
	0.1244	M_0 MLE	57.83	31.14	30.71	0.897	122.44
	0.1232	G. Pathak	48.42	23.04	17.87	0.823	70.07
	0.0954	M_t MLE	55.53	30.75	26.42	0.866	103.56
	0.1221	Petersen	43.90	21.73	19.94	0.671	78.18
(0.1, 0.025)	0.0734	M_0 CUE	49.73	21.55	17.53	0.853	68.70
	0.0720	M_0 MLE	58.43	31.54	28.07	0.897	112.02
	0.1256	G. Pathak	49.93	24.70	18.60	0.830	78.88
	0.1208	M_t MLE	54.22	25.34	20.69	0.883	81.10
	0.1420	Petersen	44.61	23.36	20.27	0.663	79.45
(0.1, 0.05)	0.0734	M_0 CUE	49.39	25.42	18.80	0.817	73.69
	0.2194	M_0 MLE	58.25	27.75	23.62	0.921	95.05
	0.0720	G. Pathak	47.40	22.15	20.30	0.805	79.59
	0.1256	M_t MLE	52.29	19.36	15.13	0.876	59.33
	0.0584	Petersen	43.96	22.45	20.68	0.674	81.08
(0.2, 0.0125)	0.2196	M_0 CUE	49.98	7.85	7.44	0.915	29.16
	0.2017	M_0 MLE	50.87	10.37	9.14	0.911	35.78
	0.1882	G. Pathak	49.52	7.92	7.63	0.914	29.86
	0.1896	M_t MLE	49.60	8.02	7.27	0.907	28.50
	0.1861	Petersen	49.39	20.24	13.41	0.681	52.55
(0.2, 0.025)	0.2196	M_0 CUE	50.20	8.87	8.37	0.912	32.80
	0.2017	M_0 MLE	50.57	9.11	8.01	0.891	31.24
	0.2132	G. Pathak	49.79	7.33	6.88	0.937	26.97
	0.1882	M_t MLE	49.26	8.01	8.09	0.914	31.73
	0.1711	Petersen	49.07	20.55	13.53	0.674	53.02
(0.2, 0.05)	0.2196	M_0 CUE	50.54	8.03	7.41	0.924	29.03
	0.1420	M_0 MLE	50.57	7.35	7.23	0.923	28.09
	0.1622	G. Pathak	50.16	6.73	6.39	0.940	25.05
	0.2017	M_t MLE	49.91	8.96	7.97	0.891	31.23
	0.1395	Petersen	49.34	18.50	13.53	0.653	48.81

Table 4.8: Simulated results for 1000 realisations from a population of size $N = 50$ with $t = 5$ and Beta-distributed capture probabilities with no plants present.

4.6.1 Results

Using the parameters given in Table 4.1, trials were simulated under model M_t with beta-distributed capture probabilities, the results of which are given in Table 4.8. The first observation from Table 4.8 is that, without plants, in many cases, the mean population estimates are a lot further from the true population size. The M_t MLE mean estimate is underestimating N by around 20-30% when $\mu = 0.05$, but then overestimates N when $\mu = 0.1$. When the mean capture probability is 0.2, it performs well in terms of the mean point population estimate.

The M_0 CUE mean population estimate is again seen to increase as the capture probability mean increases, but, as it generally underestimates when μ is small, this doesn't lead to a major positive bias. Its standard deviation estimate consistently underestimates the sample standard deviation in each trial, but the coverage proportion gets closer to 0.95 as μ increases.

The generalised Pathak estimator mean population estimate also underestimates N when $\mu = 0.05$, but has a mean population estimate of within 1% for $\mu = 0.2$. Its average estimated standard deviation also underestimates the sample standard deviation in every trial, but the coverage proportion exceeds 0.9 when $\mu = 0.2$, suggesting that the standard deviation estimate improves as μ increases.

The M_0 MLE performs well when $\mu = 0.2$ when considering its mean point population estimate, but when μ is smaller, it performs quite erratically. Its mean estimated and average standard deviations are larger than those of the other estimators (with the occasional exception of the Petersen estimator). For $\mu = 0.05$ or $\mu = 0.1$ cases, this results in it having a coverage proportion that is closer to 0.95 than that of all the other estimators. This may be a result, however, of its general overestimation of the population size.

The Petersen-type estimator performs poorly in terms of mean point population estimate when $\mu = 0.05$, where it underestimates by around 30%. When μ=0.1 its mean point population estimate improves somewhat, underestimating by just over 10%. When μ=0.2, its point population estimate is within unity of the true population size, N. Throughout the 9 trials given in Table 4.8, the sample standard deviation remain roughly the same. The mean estimated standard deviation underestimates quite significantly when $\mu = 0.2$, but otherwise is reasonably close. The coverage proportions do not ever exceed 0.7, however. This would suggest that it should not be used under model M_t.

Thus, the estimator that should be proposed under model M_t from the beta-distributed samples simulated here is the generalised Pathak estimator.

Now consider Table 4.9. The first observation is that, when the generalised Pathak estimator is considered, it has a mean point population estimate within a 1% range

N	Example p_js	t	Reps	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage	Ave. Width
400	(0.01, 0.01, 0.02, 0.03, 0.03)	5	155	M_0 CUE	361.16	208.17	214.60	0.729	841.56
				M_0 MLE	395.00	216.29	321.37	0.858	1259.75
				G. Pathak	328.01	183.36	190.50	0.671	746.77
				M_t MLE	384.92	190.47	321.14	0.839	1258.87
400	(0.1, 0.1, 0.1, 0.1, 0.01)	5	198	M_0 CUE	413.78	69.68	71.12	0.955	278.79
				M_0 MLE	425.46	78.61	76.90	0.965	301.46
				G. Pathak	399.15	67.86	67.60	0.920	264.96
				M_t MLE	394.68	72.83	68.80	0.929	269.68
100	(0.05, 0.05, 0.1, 0.15, 0.15)	5	989	M_0 CUE	106.65	35.91	29.22	0.908	114.55
				M_0 MLE	113.70	45.30	36.57	0.944	143.37
				G. Pathak	98.23	29.08	25.55	0.879	100.12
				M_t MLE	105.52	42.77	33.34	0.891	130.69
800	(0.02, 0.01, 0.03, 0.03, 0.01)	5	194	M_0 CUE	793.57	438.35	393.13	0.809	1541.08
				M_0 MLE	1189.92	791.98	910.78	0.943	3570.27
				G. Pathak	700.36	385.82	336.82	0.716	1317.52
				M_t MLE	1048.96	629.44	763.62	0.907	2993.39
400	(0.05, 0.1, 0.15, 0.15, 0.05)	5	200	M_0 CUE	427.84	61.94	59.55	0.955	233.42
				M_0 MLE	428.26	57.24	61.03	0.990	239.26
				G. Pathak	400.82	54.56	54.38	0.935	213.18
				M_t MLE	407.41	59.50	57.28	0.950	224.53
400	(0.55, 0.55, 0.50, 0.45, 0.45)	5	200	M_0 CUE	400.02	4.12	3.87	0.905	15.18
				M_0 MLE	399.29	3.73	3.88	0.965	15.23
				G. Pathak	399.80	5.31	3.90	0.955	15.35
				M_t MLE	399.71	3.83	3.88	0.940	15.22
400	(0.04, 0.05, 0.03, 0.07, 0.06)	5	200	M_0 CUE	395.95	112.53	112.49	0.880	440.95
				M_0 MLE	451.39	157.34	149.64	0.930	586.60
				G. Pathak	402.07	137.86	117.28	0.880	459.72
				M_t MLE	453.75	175.05	157.12	0.965	615.92

N	Example p_js	t	Reps	Estimator	Mean Estimate	Sample std dev.	Mean est. std dev.	Coverage	Ave. Width
400	(0.01, 0.01, 0.02, 0.03, 0.03)	5	292	M_0 CUE	331.66	177.75	190.31	0.771	746.01
				M_0 MLE	412.77	209.31	341.64	0.887	1339.21
				G. Pathak	319.28	183.33	185.84	0.685	728.51
				M_t MLE	397.68	200.89	332.01	0.860	1301.48
400	(0.5, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1)	7	100	M_0 CUE	438.08	20.41	21.37	0.590	83.38
				M_0 MLE	443.59	18.55	21.64	0.430	84.83
				G. Pathak	403.83	16.43	17.24	0.960	67.50
				M_t MLE	402.54	16.08	16.87	0.960	66.12
400	(0.6, 0.4, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)	10	100	M_0 CUE	420.79	8.42	10.05	0.440	39.39
				M_0 MLE	420.09	8.03	10.06	0.460	39.43
				G. Pathak	398.22	8.34	7.59	0.950	29.64
				M_t MLE	399.88	6.71	7.65	0.970	29.99
200	(0.3, 0.4, 0.1, 0.4, 0.3)	5	100	M_0 CUE	204.66	7.99	8.91	0.970	34.92
				M_0 MLE	203.72	7.68	8.83	0.960	34.60
				G. Pathak	199.28	8.59	8.12	0.940	31.78
				M_t MLE	197.71	8.23	8.00	0.910	30.93
400	(0.2, 0.4, 0.3, 0.1, 0.2, 0.3, 0.2)	7	100	M_0 CUE	405.38	9.95	10.65	0.950	41.74
				M_0 MLE	403.88	9.57	10.54	0.950	41.31
				G. Pathak	400.60	12.07	10.00	0.960	39.10
				M_t MLE	401.32	10.61	10.17	0.950	39.85
400	(0.6, 0.4, 0.2, 0.1, 0.1)	5	100	M_0 CUE	437.29	13.72	16.04	0.300	62.86
				M_0 MLE	436.43	13.17	16.20	0.350	63.51
				G. Pathak	399.81	10.82	11.85	0.960	46.46
				M_t MLE	400.81	11.74	12.14	0.970	47.59

Table 4.9: Average estimates under M_t based on the probabilities given in Otis et al, p. 126.

of the true population size in all but four trials. The M_t MLE has seven trials that have a mean population estimate that is biased by more than 1%. The generalised Pathak estimator also gives a reasonable point population estimate for the case in which the M_t MLE is poor, the trial with N=800. It gives a slight underestimate (of just over 5%). This is bettered still by the M_0 CUE, which gives a point population estimate of just less than 1% below the true population size, N.

The two trials where the generalised Pathak estimator has its largest mean population estimate bias correspond to trials with capture probabilities of either 0.01, 0.02 or 0.03. These are very low probabilities that would lead to few captures. This is consistent with the case when μ=0.05 in Table 4.2, where the generalised Pathak estimator also had a very low mean. In these two trials, the mean estimate of the generalised Pathak estimator is bettered by that of the M_p CUE, but this is bettered still by both MLEs.

When the mean estimated standard deviation column is considered, it can be noted that the generalised Pathak estimator tends to have a lower mean estimated standard deviation than the other standard deviation estimators considered. In some cases, this results in a coverage proportion much lower than the desired 0.95 level, but, in many cases, the coverage is at a satisfactory level. Hence, there is again evidence that the generalised Pathak estimator should be used under model M_t.

4.7 Deer Mice Pseudo Example

4.7.1 Introduction

As there are no published datasets of multiple recapture plant-capture trials, the analysis given in Amstrup et al. (2005) on deer mice (*Peronymscus maniculatus*) is extended here to a plant-capture analysis. The results, collected by V. Reid and published in Otis et al. (1978), are based on six successive nights of capturing by live-trapping and release before the next sample. The summary statistics are given in Table 4.10. The M_t MLE for this trial is $\hat{N} = 38$ with an estimated standard error of 0.62. The estimated population (s.e.) for the generalised Pathak estimator is 38 (3.87), as is that for the M_p CUE.

To convert this dataset into a pseudo plant-capture dataset the captured mice from

x	n_1	n_2	n_3	n_4	n_5	n_6
38	15	20	16	19	25	25

Table 4.10: Summary statistics for the deer mice data, giving the number of distinct animals captured as well as the number caught in sample $n_j,\ (j = 1, \ldots, t)$.

j	1	2	3	4	5
n_j	20	16	19	25	25
n_{0j}	8	8	7	12	13

Table 4.11: Summary statistics for the deer mice data, giving the number caught in sample $n_j,\ (j = 1, \ldots, t)$. n_{0j} gives the number of animals caught in sample $j,\ (j = 1, \ldots, t)$ if the 'plants' are not included. The number of distinct animals caught, x, was 23.

the first sample n_1 will be used as 'plants' and the remaining samples treated in the usual plant-capture manner. The summary statistics for this trial are given in Table 4.11. Also given in Table 4.11 are the summary statistics of the trial in which the 'plants' were not present in the population. These are denoted by $n_{0j}\ (j = 1, \ldots, t)$. The number of distinct captured animals, x, is the same for both trials, as it does not include the number of plants captured. Thus, the latter trial effectively discards the first 15 rows of the capture matrix.

4.7.2 Results

Three estimators (the M_{tp} MLE, the generalised Pathak estimator and the M_p CUE) are compared for two pseudo trials: with and without the first sample 'plants'. The results are given in Table 4.12. These show that all three plant-capture estimators give the same population estimate of 38. The M_{tp} MLE has the lowest estimated standard error of 0.83. The generalised Pathak estimator and the M_p CUE have the same form of variance estimator and give the same estimate for $x - 1$; they thus have the same estimated standard error of 3.87.

When the first trial is excluded and samples 2 to 6 used as a mark recapture trial, the M_t MLE estimates the population to be 39, with an estimated standard error of 5.10. For both the generalised Pathak estimator and the M_p CUE, the population estimate and estimated standard error are 40 and 4.12 respectively. As the analysis in the previous sections shows, on average, where the population estimates of the plant and non-plant trials are equal or nearly equal, the plant-capture estimated

Estimator	$R = 15$		$R = 0$	
	$\hat{N}$	est. s.e.	$\hat{N}$	est. s.e.
M_{tp} MLE	38	0.83	39	5.10
Generalised Pathak	38	3.87	40	4.12
M_p CUE	38	3.87	40	4.12

Table 4.12: Results of the pseudo trial. When $R = 15$ the first sample of the dataset was used as a set of planted individuals. When $R = 0$, no information from these 15 animals is used.

standard error is equal to or lower than the non-plant estimated standard error. When this is the case the population estimate is very close to the true population. In this case, the true population is not known, but the mark-recapture estimates for the full dataset are consistent with the subsets. The inclusion of plants reduces the estimated standard error for all three estimators. This is also observed in the simulation results, as the mean estimates become close to the true population.

If one relates Table 4.12 to the comparison of Table 4.2 and Table 4.8 then it appears that the capture probabilities of each sample (if they can be assumed to have been constant for each animal) of the deer mice trial may have been relatively high. This follows from noting that the population estimates of the non-plant estimators is close to the estimates when plants are included, but the former have higher estimated standard errors. This assumption of no animal heterogeneity may not hold, as the CAPTURE computer software (now part of MARK) found evidence of a trap-happy behavioural response to capture, whilst Huggins (1991) and Pledger (2000) found evidence of both behavioural and animal heterogeneity. By not taking account of heterogeneity when it is present, estimates can be negatively biased (c.f. Burnham & Overton (1978, p. 625)). This may explain why the estimates for the full model for all three estimators tested here were equal to the number of distinct animals present. Alternatively, it is possible that almost every animal present was caught in the trial.

Thus, this deer mice example has shown that the results of the simulations carried out above can apply to real data sets.

4.8 Conclusion

The strongest result from §4.5 and §4.6 is the fact the generalised Pathak estimator has been shown to be an improvement on the M_t and M_{tp} MLEs. The M_t MLE is the most commonly used estimator under model M_t, so is considered here to be the benchmark estimator.

One problem probably holding back the use of the Pathak estimator was its relative difficulty of computation. The work in this chapter, however, is evidence that modern computing power allows for it to be calculated with great accuracy. (The author recommends the use of the Java package java.math.BigDecimal for computation.)

The M_p CUE performs well in many situations, particularly when the capture probabilities are small, but its mean estimate appears to overestimate as the variance of the capture probabilities increases. This is evidence that the extra information that is gained from having knowledge of the number of captures in each sample becomes more important as the range of numbers caught increases. However, when the goal is to have as few sampling occasions as possible, the M_p CUE should be considered for use, as it is very strong in this situation.

For the reason given above, the main comparison in this chapter is between the M_{tp} MLE and the proposed estimator, the generalised Pathak estimator, both under model M_{tp}.

The strongest argument favouring the generalised Pathak estimator is the fact that its expected value does not exceed N when the capture probabilities are small. This is a key property, as the capture probabilities could be small due to the sparsity of the animal population, which could be as a result of a diminishing population. If this is the case, a cautious estimate is required rather than one that offers a false belief of a higher population abundance than really exists.

Another property of the generalised Pathak estimator is the fact that it generally has an informative limit at both ends of the confidence interval. The M_{tp} MLE can have a standard deviation estimate so high that it is larger than its mean population estimate and hence the lower end-point of the confidence interval is uninformative. This is an undesirable property, as mark-recapture and plant-capture trials can be expensive and time-consuming to run, and so one would wish to gain an informative confidence interval at the end.

To answer another question posed earlier, using plants in the process is seen to increase the stability of the estimation for each estimator. It is shown that just a relatively small number of plants can improve the estimation, and also, on occasion, cause a decrease in the average standard deviation. This was also observed in the deer mice example of §4.7.

The effect of using 10 samples rather than 5 is now discussed. The main conclusion is that there is a marked improvement in the estimation when 10 samples are used. Performing 10 samples can provide equivalent mean estimates to those trials with a mean capture probability which is double but run with 5 samples. However, in a lot of cases when $t = 5$, there is not much improvement required for the point estimates of N, as many of the mean population estimates are very close to the true population size. Thus, the recommendation here is that only five samples should be run, unless the practitioner has a strong sense that the capture probabilities are below 0.1 on average.

As the focus here was on models M_t and M_{tp}, there has been no analysis of model selection. Consequently, the results here only apply to these models. If one wishes to use the above conclusions, they should firstly carry out a model selection procedure to confirm that a time-heterogeneous only model can be assumed. A test of model M_t against the more general Jolly-Seber model is given in Stanley & Burnham (1999). Other frequently used model selection procedures use the *Akaike's Information Criterion, AIC* or the *Bayesian Information Criterion, BIC*. Both of these methods are excellently described and compared in Buckland et al. (1997).

Chapter 5

MARK-RECAPTURE AND PLANT-CAPTURE ESTIMATION WITH SPARSE DATA

5.1 Introduction

This chapter is concerned with point and interval estimation of a closed population of size N under model M_t of the Otis-class, and the equivalent plant-capture model, M_{tp}. The capture histories are given by an $(X+R)\times t$ matrix, D (see §1.5), where, without loss of generality, the first X rows give the capture histories for the distinct target animals caught, the next R rows give the capture histories for the plants and t is the number of samples. The stipulation for D here is that it should be sparsely filled with 1s, representing captures. By definition, each of the first X rows must have at least one capture, but there should at most be few recaptures for it to be defined as a "sparse data" set. Sparseness is difficult to define in more quantitative terms, since it must be defined in terms that are known, rather than the unknown N. This is why it is difficult to get an accurate mathematical definition. The condition that produces a sparse data set is a set of small capture probabilities, $p_j, j = 1, \ldots, t$.

Most population estimators perform poorly in such situations, as shown in the previous chapters and by other authors (Chao (1989), Chapman (1951), Otis et al. (1978, p. 26)). These biases are generally caused by the estimators underestimating. Darroch's (1958) MLE, however, has the problem that it gives an infinite estimate of population size when the total number of captures, z, equals x, which becomes increasingly probable as the capture probabilities decrease.

Another problem caused by having small sample sizes is that of model selection, as detailed in Borchers et al. (2002, §6.7.3). Although model selection will not be considered further here, it illustrates that sparse data sets can be more problematic than the larger sample size trials.

Various authors have sought to improve the estimation of sparse data sets; Chao (1989) looked at the problem under models M_t and M_h and Gazey & Staley (1986) sought to find a Bayesian solution to sparse data problems. The recommended estimator by Chao (1989) under M_t is one that estimates the number of animals never caught, using the numbers of animals caught exactly once and exactly twice. She shows this to be a bias-corrected form of an estimator that reduces to the Petersen estimator when $t = 2$. She also states that this estimator has a bias of order $O\left(\frac{1}{N}\right)$, compared to a bias of order $O(1)$ for the M_t MLE as calculated by Darroch (1958). As a result, she recommends the use of the former rather than the latter.

Chao's (1989) estimator is also recommended by Wilson & Collins (1992), who compare it with the estimators of Darroch & Ratcliff (1980) and Zelterman (1988). They conclude that "It is found that the bias adjusted estimator of Chao (1989) is the best to use when the number of captures is relatively small . . . ".

We will compare here Chao's (1989) estimator under model M_t with the model M_0 conditionally unbiased estimator, derived by Goudie & Ashbridge (2005). As was shown in the previous chapters, the M_0 CUE performed very well in terms of mean point estimate when the capture probabilities were low. Thus, it is tested here against an estimator designed for sparse data trials.

It was also shown in the previous chapters that the generalised Pathak estimator performed as well as, if not better than, the M_0 CUE in many occasions with small capture probabilities. However, the combination of a large population size and a large number of samples precludes its use here, as the computations proved to be too difficult.

This chapter also looks at the effect that plants have in improving estimation. Yip (1996), in his work on the continuous-time heterogeneous model, states that "The effect of R is more significant when p [the capture probability] is small". He does not, however, consider capture probabilities that would lead to the "sparse data" case, which will be considered here. For model M_p Ashbridge & Goudie (2009) derive a generalised version of the M_0 CUE, which uses a ratio of Gould-Hopper numbers, (Gould & Hopper 1962), in place of the C-numbers.

We also test here whether including plants can allow the experimenter to take fewer samples whilst getting almost unbiased estimates of population size. Goudie (1995) concludes that, under a continuous-time framework, ". . . the use of plants can provide a useful reduction in the average time taken to achieve complete coverage . . . ". Through simulation, work is done to consider whether the number of samples taken can be reduced by including plants.

This chapter proceeds as follows: §5.2 provides a reminder of the plant-capture CUE of Ashbridge & Goudie (2009) and gives a generalisation of the estimator of

Chao (1989); §5.3 gives simulation results and plots detailing the effect of plants for various scenarios leading to sparse matrices. Finally, the conclusions of this work are given in §5.4.

5.2 Theory

The sparse data estimator proposed by Chao (1989) for model M_t is here generalised to model M_{tp}, the equivalent plant-capture model. For this estimator some additional notation is required, which is given below.

X_R = The number of distinct animals from the plant population caught in the trial.

f_k = The number of animals from both target and plant populations caught exactly k times, $k = 0, 1, \ldots, t$.

Z_j = The number of animals from both populations captured only in the j^{th} sample, $j = 1, \ldots, t$.

When $R = 0$ the justification for Chao's (1989) M_t estimator is based on the equation $N = X + f_0$. To generalise this estimator to model M_{tp} we estimate the augmented population and then subtract the known number of plants, giving $N = (X + X_R - R) + f_0$. In both cases, since f_0 is not observed, the final term must be estimated. This requires the two observed statistics f_1 and f_2, the number of animals caught exactly once and twice respectively in the t samples. Thus, under model M_{tp} we have

$$\begin{aligned}
E[f_0] &= (N+R)\prod_{j=1}^{t}(1-p_j) \\
E[f_1] &= (N+R)\prod_{j=1}^{t}(1-p_j)\left[\sum_{k=1}^{t}\frac{p_k}{1-p_k}\right] \\
E[f_2] &= (N+R)\prod_{j=1}^{t}(1-p_j)\left[\sum_{k=1}^{t}\sum_{l=k+1}^{t}\frac{p_k p_l}{(1-p_k)(1-p_l)}\right].
\end{aligned}$$

Combining the above, we get

$$\begin{aligned}
\Big\{E[f_1]\Big\}^2 - 2E[f_0]E[f_2] &= (N+R)^2\prod_{j=1}^{t}(1-p_j)^2\left[\sum_{k=1}^{t}\left(\frac{p_k}{1-p_k}\right)^2\right] \\
&= \sum_{j=1}^{t}\Big\{E[Z_j]\Big\}^2.
\end{aligned}$$

Thus, rearranging to obtain an expression for the expected value of the unknown f_0, we get

$$E[f_0] = \frac{\{E[f_1]\}^2 - \sum_{j=1}^t \{E[Z_j]\}^2}{2E[f_2]},$$

which leads to an estimate (adjusted to allow for the possibility of $f_2 = 0$) of N:

$$\hat{N}_t = (x + x_R - R) + \frac{f_1^2 - \sum_{j=1}^t Z_j^2}{2(f_2+1)}. \tag{5.1}$$

A variance estimator for (5.1) (for the $R = 0$ case) is also given in Chao (1989) but it is stated that it "slightly underestimates" the sample standard error.
Also stated in the conclusion of Chao (1989) is:

> "For sparse data, the proposed $\hat{N}_C$ [(5.1)] is preferable to Darroch's (1958) MLE in the sense of having smaller bias as well as smaller variance. However, when data are not sparse so that there are relatively more recaptures, Darroch's MLE would perform better than the proposed $\hat{N}_C$, for in such cases Darroch's MLE will have negligible bias and smaller variance."

It is this split in optimality that makes this awkward, since it is difficult in practice to know whether the data are sparse or not. Without knowing the true population size, and even when one does, there is no clear boundary between sparse data and non-sparse data. For this reason, a more unified approach is sought.
Chao's estimator, (5.1), is compared here with the conditionally unbiased estimator, CUE, under model M_0 of Goudie & Ashbridge (2005), which is the unique unbiased estimator of N under the conditional distribution given $Z = z$ in the case where $N \leq z$. This estimator is generalised to model M_p in Ashbridge & Goudie (2009), and was given in §2.3 as

$$\tilde{N}_c = x + \frac{G(z, x-1, t, Rt)}{G(z, x, t, Rt)}, \tag{5.2}$$

where $G(z, x, t, Rt)$ is a Gould-Hopper number (Gould & Hopper 1962) defined by (2.19) and given again here for convenience:

$$\begin{aligned} G(z,x,t,Rt) &= \frac{1}{x!}\Delta^x \left[(Rt+\omega t)_z\right]_{\omega=0} \\ &= \frac{z!}{x!}\sum_{k=0}^{x}(-1)^k \binom{x}{k}\binom{Rt+xt-kt}{z}. \end{aligned}$$

As with the previous chapters, what is actually used in the simulation results is $\hat{N}_U$ where this is the integer rounded value (c.f. p.45).

From §4.4.4 and defining N_{c-1} to be (5.2) with $(x-1)$ distinct captures, we get:

$$\text{var}\left(\tilde{N}_c\right) = \begin{cases} \left(\tilde{N}_c - x\right)\left(\tilde{N}_c - \tilde{N}_{c-1}\right) & x = \frac{z}{t} - R + 1, \ldots, z; \\ \\ 0 & x = \frac{z}{t} - R. \end{cases} \tag{5.3}$$

Also given in Ashbridge & Goudie (2009) is the probability distribution for (Z, X) under model M_p. This distribution has probability function

$$p(z,x) = \frac{(N)_x}{z!} G(z,x,t,Rt) p^z (1-p)^{Nt+Rt-z} \tag{5.4}$$

for $z = 0, \ldots, Nt + Rt$, $x = \min(1,z), \ldots, \min(N,z)$. Under M_p one can calculate the expected value for the M_p CUE by summing the product of (5.2) and (5.4) over all Z and X. This is used in the results section to measure the simulation error in the CUE estimates.

5.3 Results

This section begins in an analogous way to Chao's (1989) paper, using 40 trials and the same sets of probabilities that she gives in her Table 1, (Chao 1989, p. 433), given here in Table 5.1. These scenarios are given in Table 5.1 for the M_{tp} trials, with the M_p adjusted scenarios given in Table 5.2. The results therein are recalculated and given in Table 5.3.

The analysis then moves on to examine the effect that the number of samples plays in the quality of the estimators, as the 40 samples that are used by Chao (1989) would be very time-consuming and expensive for the practitioner. Thus, a modified set of Trials, for the case when there are 10 sampling occasions instead of 40, is given in Table 5.2. Where possible, the number of samples with a particular capture probability was directly scaled down by a factor of 4, but when this was not possible the sample probabilities were chosen to give 'more sparse' scenarios than the $t = 40$ case. Also, when the M_{tp} scenarios were approximated by M_p scenarios (for which the exact expected value of the M_p CUE can be calculated), the average capture probability over all samples was used, given in the final column of Table 5.2.

Finally, the trials are simulated with the inclusion of plants, to test whether including plants can improve estimation.

Trial	Sample	p_j
1	$j = 1, 20$	0.003
	$j = 21, 40$	0.005
2	$j = 1, 40$	0.005
3	$j = 1, 20$	0.003
	$j = 21, 30$	0.005
	$j = 31, 40$	0.01
4	$j = 1, 30$	0.005
	$j = 31, 40$	0.01
5	$j = 1, 20$	0.005
	$j = 21, 40$	0.01
6	$j = 1, 40$	0.01

Table 5.1: Model scenarios for $t = 40$ as given in Chao (1989), which lead to sparse data situations.

Trial	Sample	p_j	M_p probability
1	$j = 1, 5$	0.003	$p = 0.004$
	$j = 6, 10$	0.005	
2	$j = 1, 10$	0.005	$p = 0.005$
3	$j = 1, 6$	0.003	$p = 0.0048$
	$j = 7, 8$	0.005	
	$j = 9, 10$	0.01	
4	$j = 1, 8$	0.005	$p = 0.006$
	$j = 9, 10$	0.01	
5	$j = 1, 5$	0.005	$p = 0.0075$
	$j = 6, 10$	0.01	
6	$j = 1, 10$	0.01	$p = 0.01$

Table 5.2: Model scenarios for $t = 10$, similar to those given in Chao (1989), which lead to sparse data situations. The final column gives the constant capture probability used for the M_p CUE estimates.

5.3.1 Comparison of the Chao and CUE estimators under sparse data conditions

For the trials specified in Table 5.1, simulated results, based on 500 realisations for each trial, are given in Table 5.3. The results in Table 5.3 for the estimator $\hat{N}_t$ are similar to those of Chao (1989, p. 433, Table 1), but are based on new simulations. The author believes that the discrepancies between the values in Table 5.3 and Chao's Table 1 can be attributed to simulation error.

It can be seen from Table 5.3 that there is not much difference between the simulated results of Chao's estimator and the M_0 CUE in terms of bias. The M_0 CUE has a lower bias in 11 of the 18 cases simulated, but the margin of improvement is only small in each case. This shows that there is no significant loss of accuracy when using the CUE rather than Chao's estimator. The difference in the quality of the

Trial	N	Chao $\hat{N}_t$			M_0 CUE		
		Mean	Sample s.d.	Chao s.d.	Mean	Sample s.d.	CUE s.d.
1	250	234.47	135.28	95.81	239.47	131.76	152.43
	500	496.35	219.15	102.83	497.49	202.82	218.89
	1000	996.77	293.99	147.36	1001.56	275.52	294.60
2	250	243.08	126.13	70.23	245.50	119.76	130.03
	500	497.97	188.95	91.02	499.34	161.86	168.52
	1000	999.77	245.37	130.31	995.05	224.08	228.29
3	250	253.32	151.46	77.43	256.87	145.09	146.40
	500	499.57	179.31	89.04	499.75	149.66	158.74
	1000	1011.72	251.31	128.29	1013.47	219.58	222.50
4	250	255.28	134.77	63.66	255.65	126.25	120.28
	500	505.36	168.04	81.42	503.98	134.12	133.25
	1000	1003.98	196.47	114.90	998.18	172.30	172.30
5	250	252.89	98.38	51.60	255.32	89.41	86.13
	500	500.50	120.57	72.19	502.90	105.55	106.73
	1000	1003.90	168.87	102.86	996.35	145.14	145.34
6	250	251.60	77.45	42.82	249.45	50.95	54.86
	500	503.23	89.35	60.58	504.10	76.40	77.34
	1000	1002.70	130.61	85.63	998.25	106.03	106.20

Table 5.3: Comparison of the Chao and M_p CUE estimators under model scenarios with $t = 40$ and $R = 0$ as given in Chao (1989), which lead to sparse data situations.

Trial	N	Chao $\hat{N}_t$			M_0 CUE		
		Mean	Sample s.d.	Chao s.d.	Mean	Sample s.d.	CUE s.d.
1	250	50.00	29.42	45.88	49.98	29.40	37.93
	500	163.84	79.44	141.74	165.27	78.74	121.54
	1000	538.49	260.04	421.85	545.57	258.79	403.96
2	250	71.67	40.77	64.85	71.57	40.39	54.15
	500	227.53	116.92	182.68	228.97	115.10	169.42
	1000	710.82	383.34	487.12	716.81	379.21	534.13
3	250	64.55	35.87	57.89	66.37	36.92	49.83
	500	210.02	104.30	175.92	218.69	106.36	161.80
	1000	655.92	332.07	462.48	682.19	339.66	503.19
4	250	91.96	50.63	80.01	93.26	50.67	69.81
	500	288.95	152.89	218.82	294.55	151.01	217.82
	1000	836.72	480.52	485.17	853.90	482.57	630.51
5	250	126.46	70.70	103.69	128.86	70.54	96.37
	500	372.23	207.25	245.29	380.48	204.89	278.73
	1000	953.39	594.22	403.84	966.33	584.58	684.53
6	250	178.54	99.87	127.44	180.83	96.36	131.48
	500	442.74	261.98	198.60	449.02	255.03	308.78
	1000	984.84	548.31	234.91	987.72	531.23	565.35

Table 5.4: Comparison of the Chao and M_p CUE estimators under model scenarios with $t = 10$ and $R = 0$ as given in Chao (1989), which lead to sparse data situations.

standard deviation estimates is more pronounced. It is evident that the estimator of the standard deviation given by the square root of (5.3) for the M_0 CUE estimates the appropriate sample standard deviation more closely than the standard deviation estimator given by Chao estimates the sample standard deviation of her estimator. The estimated standard deviation for the Chao estimator is between 29% and 53% smaller than the sample standard deviation in each case, which would lead to too narrow a confidence interval. The corresponding difference between the estimated and sample standard deviations for the CUE is between -16% and 5%, which leads to wider, but more realistic, confidence intervals.

In every trial given in Table 5.4, however, the means of both estimators underestimate the size of the true population. With the exception of the first and fourth rows, the M_0 CUE is consistently higher than $\hat{N}_t$, albeit only marginally. One might be tempted to conclude that, in order to get estimates close to the true population when the capture probabilities are small, one must sample on many occasions. This becomes very expensive and time-consuming, and also makes the assumption of a closed population increasingly unjustifiable. It seemed plausible that the strong negative bias evident when there are few samples might be reduced by introducing plants into the population before sampling begins. In the following section the effect of introducing plants on the estimators' biases are considered, by means of simulation.

5.3.2 Comparison of plant and non-plant estimation when t=40

In order to test whether the use of plants improves the estimation of population size, for each integer value of R in $[0, 200]$, 1000 realisations of the sampling process are simulated for the case where $N = 500$ and $t = 40$. In Figure 5.1, for each value of R shown on the horizontal axis, the mean of the 1000 realisations' estimates is plotted for both Chao's $\hat{N}_t$ and the CUE $\hat{N}_U$. This is done for each Trial given in Table 5.1. Figure 5.2 contains plots of the sample standard deviations from both estimators for each of these Trials.

It is evident from Table 5.3 that the mean point estimate for the no plant cases, where $R = 0$, is almost unbiased for both estimators for all Trials given in Table 5.1, so there is not much room for improvement. It can be seen from Figure 5.1 that the inclusion of any number of plants has no effect on the point population estimation, as all six plots oscillate around the true value of $N = 500$ with no obvious convergence. Figure 5.2, however, shows that, for the M_p CUE, there is a reduced sample standard deviation as more plants are added. The reduction in sample standard deviation between no plants and 50 plants can be as much as 25%. However, the reduction in size becomes negligible as the number of plants increase above

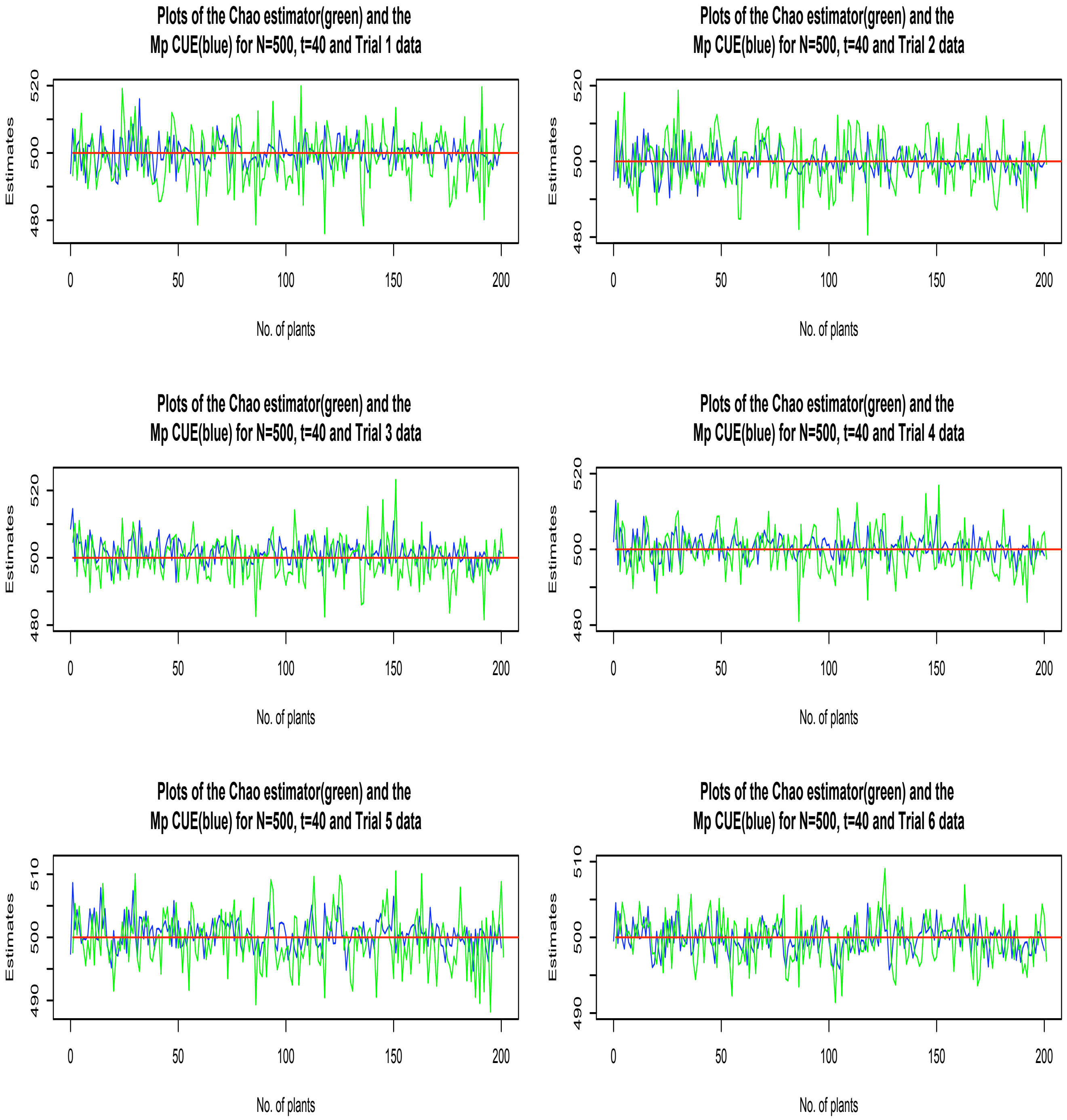


Figure 5.1: Plots of means of the simulated (blue) and expected (red) M_p CUE and Chao (green) estimators against the number of plants included for the case $N = 500$ and $t = 40$.

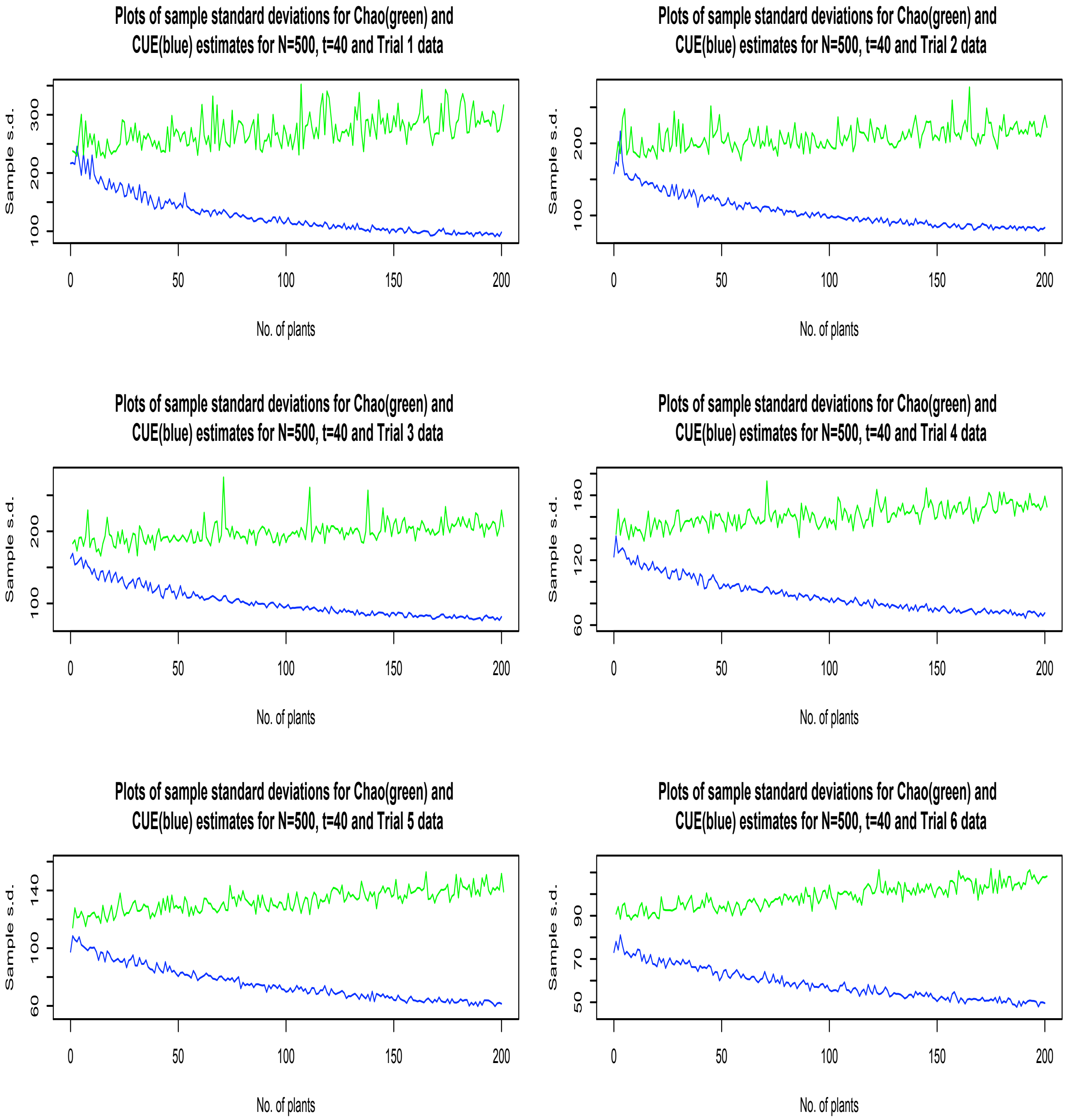


Figure 5.2: Plots of the sample standard deviations from simulated (blue) M_p CUE and Chao (green) estimates against the number of plants included for the case $N = 500$ and $t = 40$.

50. This is consistent with the conclusion of Yip (1996, p. 2037), which states that adding too many plants should be avoided, as "...the improvement is quite insignificant ...".

As for the sample standard deviation of the Chao estimator, no reduction in size is evident as the number of plants increase. In fact, the sample standard deviation appears to increase linearly as R, and hence the augmented population, increases. The suggestion here would be that the modified Chao estimator is not utilising the information gained from the plant captures (*i.e.* not improving the estimate of f_0) and so the spread of estimates increases as the augmented population increases.

Here, we conclude that plants offer no increase in either estimate's bias but equally should not be considered an improvement if the non-plant estimate is almost unbiased. For the M_p CUE introducing plants may reduce the standard deviation of any estimate but no such reduction would be expected from any standard deviation estimator for the Chao estimator.

5.3.3 Comparison between 40 samples and 10 samples

A more pertinent question is whether the use of plants is more efficacious when fewer samples are taken and the data are consequently more sparse. Given in Figures 5.3 – 5.6 are plots of the means of the estimates based on 500 realisations of the trials given in Table 5.2, but with the inclusion of between 0 and 200 plants.

Figure 5.3 gives the means of the estimates for the simulated CUE $\hat{N}_U$ estimator and the Chao $\hat{N}_t$ estimator when the true population size is 100 and the number of plants included ranges from 0 to 200 in increments of 1. Also included in the plots is the exact mean of the CUE, where this assumes a constant capture probability, p, in each sample. For Trials 2 and 6, this estimate should be the asymptotic limit of the simulated mean, as the number of replicates goes to infinity. For the other trials, it should offer an approximate asymptotic limit. Also marked in black on the plots is the true population size of 100, for comparison.

Similar plots are given in Figures 5.4, 5.5 and 5.6 for populations $N = 250$, 500 and 1000 respectively. The ensuing analysis is concerned with answering the following questions: what is the effect of the true population size on the accuracy of the estimation and what is the effect of including plants on the quality of estimation?

To answer the first of these questions, it is evident that a similar story emerges from each of Figures 5.3 – 5.6. With the exception of Trials 5 and 6 when $N = 1000$, the estimates from all estimators when no plants are included are severely nega-

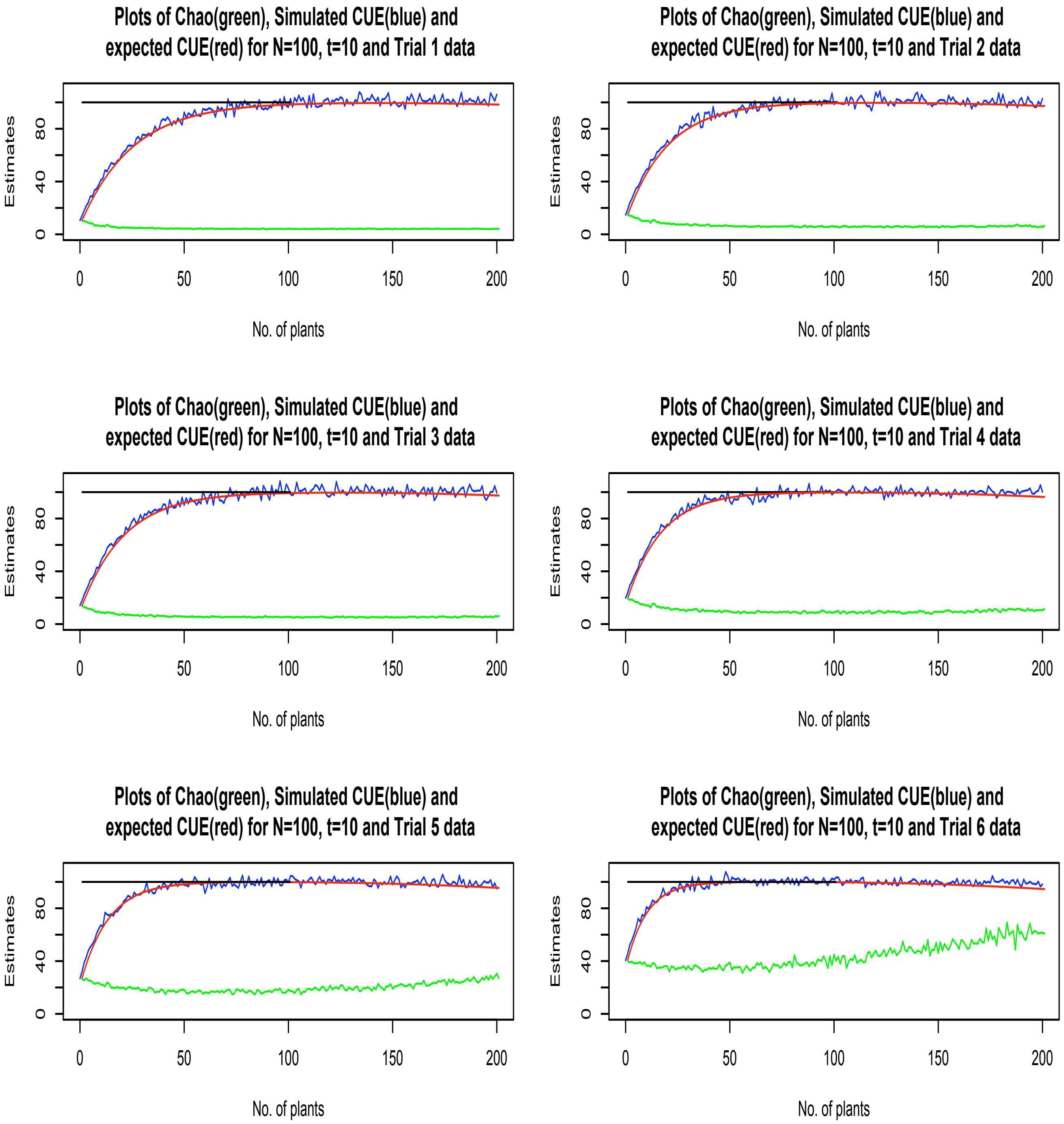


Figure 5.3: Plots of means of the simulated (blue) and expected (red) M_0 CUE and Chao (green) estimators against the number of plants included for the case $N = 100$ and $t = 10$.

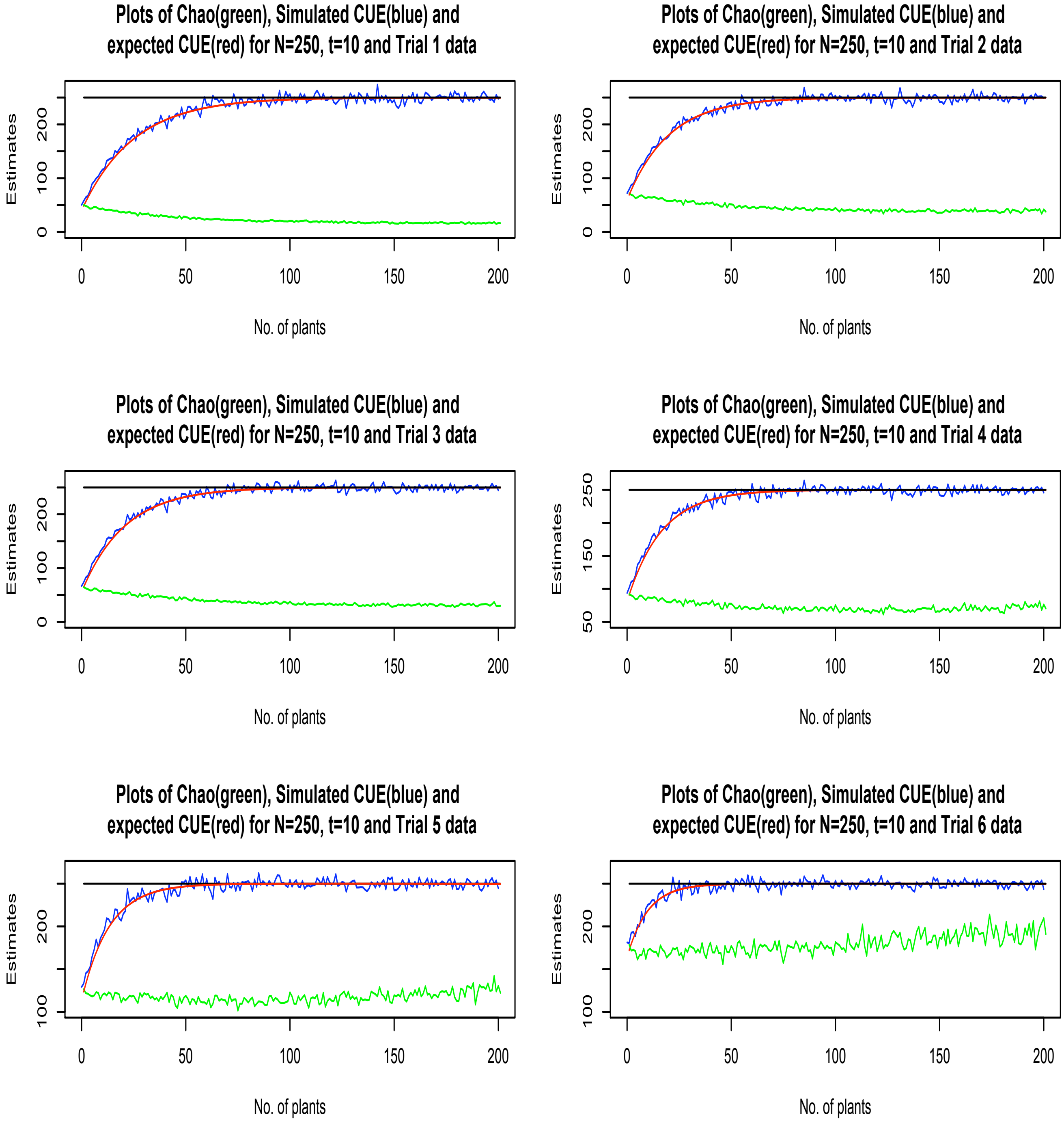


Figure 5.4: Plots of means of the simulated (blue) and expected (red) M_0 CUE and Chao (green) estimators against the number of plants included for the case $N = 250$ and $t = 10$.

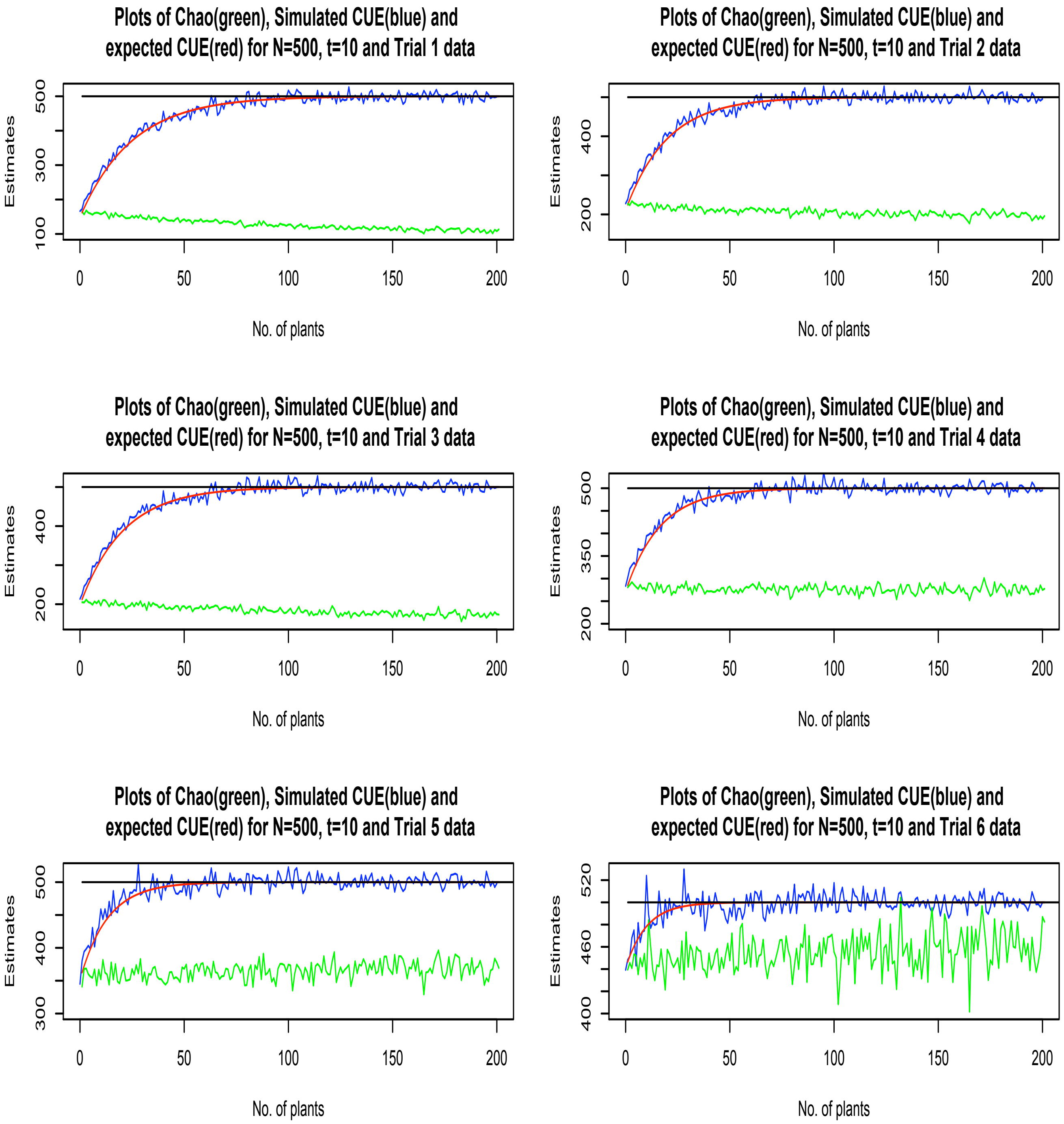


Figure 5.5: Plots of means of the simulated (blue) and expected (red) M_0 CUE and Chao (green) estimators against the number of plants included for the case $N = 500$ and $t = 10$.

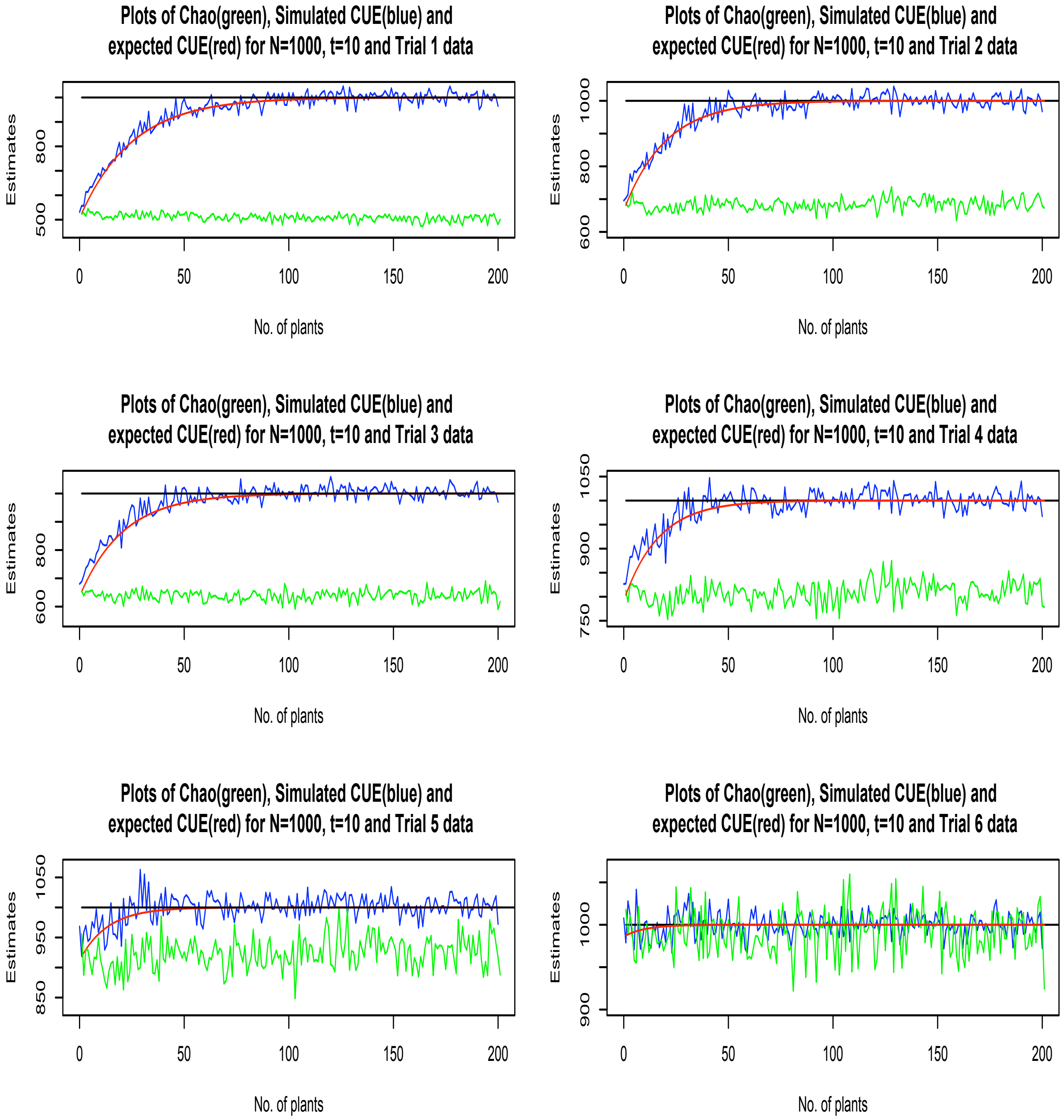


Figure 5.6: Plots of means of the simulated (blue) and expected (red) M_0 CUE and Chao (green) estimators against the number of plants included for the case $N = 1000$ and $t = 10$.

tively biased. Thus, for Trials 1 – 4 of Table 5.2 it appears that the total population size does not affect the relative bias of the estimator, whether it be Chao's or the CUE. As the population reaches 1000, however, it is noted that for the 'less sparse' cases, Trials 5 and 6, the estimate when $R = 0$ from each estimator is only mildly negatively biased, with the values given in Table 5.4.

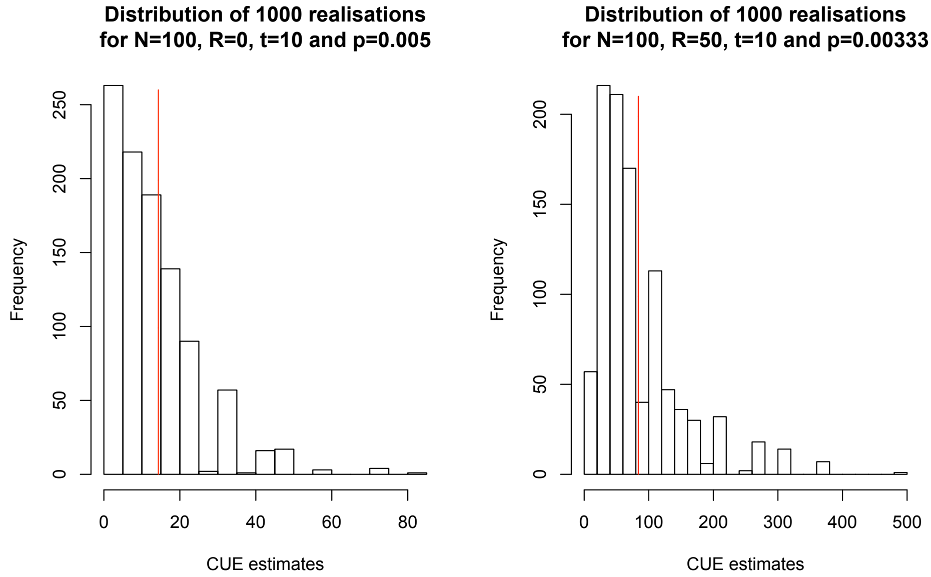
Regarding the more important question of whether plants decrease the bias of the estimators, the answer depends on the estimator. When one considers Chao's estimator it can be seen from most plots that the inclusion of plants increases the bias of the estimate, and it appears that the plants can have a negative effect for this estimator. Since the Chao estimator effectively estimates the size of the augmented population, it appears that the negative bias is so great that the subtraction of the R term can make the estimate negative, and so the estimates given in Figures 5.3 – 5.6 are bounded below by X, which is the smallest size that the population can be. For the CUE the simulated means of the estimator (shown in blue on the plots) can be seen to approximate the expected value (given in red on the plots) very closely, and the oscillations in the simulated line can be attributed to random variation. It is also evident that, in every trial and for every population size considered, the inclusion of a sufficient number of plants leads to an unbiased estimate of the true population size. When one wishes to find a balance between obtaining an unbiased estimate and using as few plants as possible, the optimal number of plants to include is not clear, but in most cases $R = 50$ appears to be sufficient to give an almost unbiased estimate.

5.3.4 Distribution of the M_p CUE estimates with and without plants

Given in Figures 5.7 and 5.8 are histograms of the distribution of 1000 realisations of the M_p CUE under Trial 2 probabilities. In Figure 5.7 $N = 100$ and in Figure 5.8 $N = 250$. In both Figures 5.7 and 5.8 the left hand plot gives the distribution of estimates when no plants are present and the right hand plot gives the distribution of estimates when $R = 50$ (with a slightly reduced constant capture probability to give the same expected number of captures). The histograms can be thought of as slices of the Trial 2 (top right) plots in Figures 5.3 and 5.4, sliced at $R = 0$ and $R = 50$.

What both Figures 5.7 and 5.8 both show is that, when $R = 0$, there is not a single realisation that exceeds the true population size, N and the mode of both non-plant histograms is far below N. Introducing plants is seen to shift the mean and mode up towards N, and also stretch out the upper tail.

The addition of plants can cause a few outliers that are more than four times the true population, which will have some influence on shifting the mean upwards. The density of estimates that are close to N is much increased. Thus, despite the fact that underestimation is preferable to overestimation in most ecological circumstances, the mean improvement gained by inserting plants leads to the recommendation of their use.


Distribution of 1000 realisations for N=100, R=0, t=10 and p=0.005

Distribution of 1000 realisations for N=100, R=50, t=10 and p=0.00333

Figure 5.7: Distribution of 1000 realisations of M_p CUE estimates for the case $N = 100$ and $t = 10$ with $R = 0$ (left) and $R = 50$, with the mean estimate given by the red vertical line.

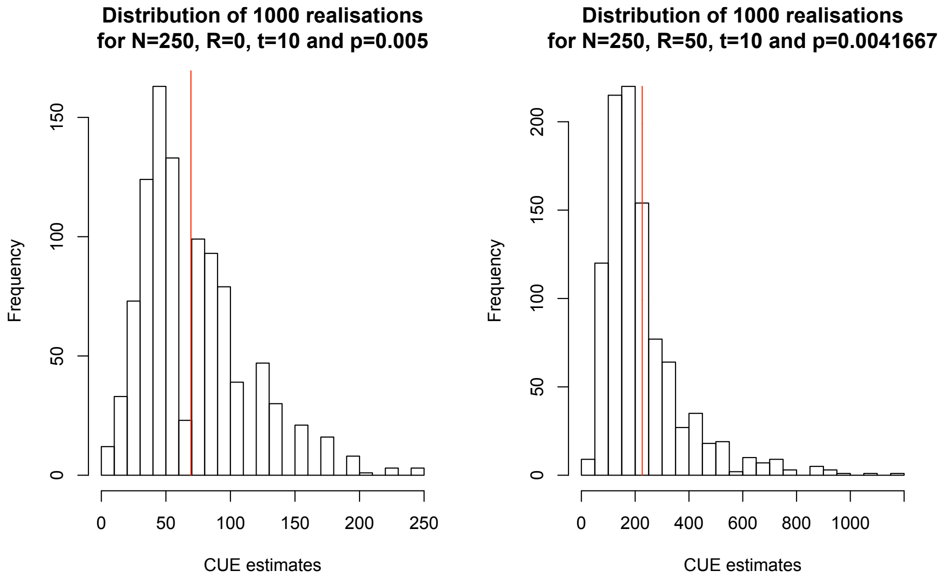


Figure 5.8: Distribution of 1000 realisations of M_p CUE estimates for the case $N = 250$ and $t = 10$ with $R = 0$ (left) and $R = 50$, with the mean estimate given by the red vertical line.

5.4 Conclusion

This chapter has shown that the M_0 CUE of Goudie & Ashbridge (2005), later generalised to model M_p in Ashbridge & Goudie (2009), is preferable to the estimator of Chao (1989) under sparse data conditions. The CUE has a bias equal to, if not lower than, Chao's estimator, as well as having a superior variance estimator. It was also shown that the mean of the M_p CUE is effectively unbiased when a sufficient number of plants (at least 50) are inserted into the target population, whereas Chao's estimator is not improved by the inclusion of plants. This should not be seen as a problem with plant-capture, but rather a problem specific to the Chao estimator. It is possible that an improvement to this estimator can be made so that, with the aid of the planted individuals, f_0 is estimated more accurately.

Thus, it is recommended that the number of samples carried out in mark-recapture experiments, when the capture probabilities are very small, can be vastly reduced with the inclusion of around 50 plants. It has been shown here that one can expect almost unbiased results when using the M_p CUE with the inclusion of planted individuals. This would drastically cut down on the number of samples, which would make it more plausible that the closure assumption, a requirement of model M_{tp}, holds.

Chapter 6

COVERAGE ESTIMATORS

6.1 Introduction

The concept of coverage is important when one wants to consider which estimator is optimal when estimating the size of animal populations. For this reason, an alternative estimate of coverage is proposed here.

Chao et al. (1992), offer three estimators of coverage under model M_{th}, which can be used under the special case of model M_t, in which the probability of capture in any sample is the same for all animals. These estimators are given in §6.2.

Estimating coverage under the homogeneous models can equivalently be thought of as estimating the inverse population, which, when multiplied by x, gives a coverage estimate. An estimator of $\frac{1}{N}$ that exists in the literature is given by Pathak (1964), which is shown to be the minimum variance unbiased estimator under model M_f. This estimator is generalised to model M_{fp} below in §6.3.

A modified version of this generalised Pathak estimator to give the optimised estimator under model M_p, using the Rao-Blackwell Theorem, is presented in §6.4. These estimators are compared via simulation, and the results are presented in §6.5.

6.2 Chao coverage estimators

Three estimators that directly estimate coverage (as opposed to estimators of $\frac{1}{N}$) have been proposed by Chao et al. (1992) for Model M_{th}. These estimators are used here under Model M_t, which is a special case of M_{th}, assuming homogeneity of animals in each sample.

For the derivation of the estimators, however, it is assumed that each animal has an unknown probability of capture, p_i, $i = 1, \ldots, N$, the heterogeneity component. The sample coverage is derived by summing the probabilities for all captured animals, and dividing this by the sum of all probabilities. This, however, is equivalent to 1 minus the ratio of the sum of capture probabilities of those animals never caught to the total sum. Thus, we get:

$$\hat{C} = 1 - \frac{\sum_{i=1}^{N} p_i I \left[\text{the } i^{th} \text{ animal never captured}\right]}{\sum_{i=1}^{N} p_i}. \tag{6.1}$$

If there is also an unknown time-effect, $e_j,\ j = 1, \ldots, t$, in each sample, we get:

$$E[\hat{C}] = 1 - \frac{\sum_{i=1}^{N} p_i \prod_{j=1}^{t} (1 - p_i e_j)}{\sum_{i=1}^{N} p_i}$$

Multiplying top and bottom with $\sum_{j=1}^{t} e_j$ and using the fact that finite sums are interchangeable gives

$$\begin{aligned} E[\hat{C}] &= 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{t} \left[p_i e_j \prod_{s=1}^{t} (1 - p_i e_s) \right]}{\sum_{i=1}^{N} \sum_{j=1}^{t} p_i e_j} && (6.2) \\ &\approx 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{t} \left[p_i e_j \prod_{s \neq j} (1 - p_i e_s) \right]}{\sum_{i=1}^{N} \sum_{j=1}^{t} p_i e_j}. && (6.3) \end{aligned}$$

Note that the approximation (6.3) of (6.2) should be an underestimate of $\hat{C}$, since factors in the product that are less than unity are being removed from the product, thus increasing the overall sum.

Observing that the numerator of (6.3) is just $E[f_1]$, Chao et al. (1992, p. 205) derived her first estimator of coverage, given by (6.4). She then gave two bias-corrected estimators, given by (6.5) and (6.6). All three estimators use the frequency of captures, f_k, as explanatory variables, and are as follows:

$$\begin{aligned}
\hat{C}_1 &= 1 - \frac{f_1}{\sum_{k=1}^{t} k f_k} \qquad (6.4)\\
\hat{C}_2 &= \min\left\{1, 1 - \frac{f_1 - 2f_2/(t-1)}{\sum_{k=1}^{t} k f_k}\right\} \qquad (6.5)\\
\hat{C}_3 &= \min\left\{1, 1 - \frac{f_1 - 2f_2/(t-1) + 6f_3/[(t-1)(t-2)]}{\sum_{k=1}^{t} k f_k}\right\}. \qquad (6.6)
\end{aligned}$$

Note that (6.5) and (6.6) differ from Chao et al. (1992) in that they are bounded above by unity, since they do not exclude the probability of estimates greater than unity for the coverage. Note also that the denominator in these equations is simply z, the total number of animals captured. These three estimators were proposed for Model M_{th} scenarios, but are used in this thesis for M_t models only.

6.3 Pathak's inverse population estimator under M_{fp}

Pathak (1964) wrote his paper with the purpose of estimating population size and its inverse when sampling was carried out under model M_f, the fixed sample size model. This estimator is generalised here to model M_{fp}. As seen in Chapter 2, eq. (2.1), the sufficient statistic under model M_{fp} is X, the number of distinct animals caught.

Pathak (1964) begins his paper by making use of the inclusion-exclusion principle (c.f. Johnson et al. (2005, p. 432)):

Lemma 1.1: Let $A_1, \ldots, A_x$ be x events defined on a probability space. Let $A = \bigcup_{i=1}^{x} A_i$ and $B_i = (A - A_i)$, $i = 1, \ldots, x$. Then

$$p\left[\bigcap_{i=1}^{x} A_i\right] = p(A) - \overset{\bullet}{\sum} p(B_1) + \overset{\bullet}{\sum} p(B_1 \cap B_2) - \ldots, \qquad (6.7)$$

where $\overset{\bullet}{\sum}$ is taken over all combinations of Bs chosen from $B_1, \ldots, B_x$.

□

To generalise Pathak (1964) to get an estimate of $\frac{1}{N}$ under model M_{fp} we use equation (2.6), as given in §2.2, which is restated below:

$$p(X = x) = \frac{(N)_x a(x, \mathbf{n}, R)}{A(N, \mathbf{n}, R)}, \qquad x = 0, \ldots, N. \qquad (6.8)$$

To get an unbiased estimate of the population inverse we firstly assume that at least one animal from the target population is captured in each of the first two samples. This assumption holds with probability 1 if both n_1 and n_2 are chosen to be greater than R. If we let u_{11} and u_{21} be the first units from the target population in samples 1 and 2 respectively, then $p(u_{11} = u_{21}) = \frac{1}{N}$. Thus, we have an unbiased estimator of $\frac{1}{N}$, given by:

$$t_1 = \begin{cases} 1 & \text{if } u_{11} = u_{21} \\ 0 & \text{otherwise} \end{cases}. \tag{6.9}$$

Since X is a complete sufficient statistic under model M_{fp}, we can use the Rao-Blackwell Theorem (6.9) to get the minimum variance unbiased estimate of $\frac{1}{N}$:

$$\hat{N}_{-1}(x) = E\left[t_1 \mid X = x\right] = \frac{p\left(u_{11} = u_{21} \bigcap X = x\right)}{p\left(X = x\right)}. \tag{6.10}$$

To find the numerator in terms of $x, n_1, \ldots, n_t$, we let $u_{(1)}, \ldots, u_{(x)}$ represent the x distinct animals captured in the sample. Thus, if we use Lemma 1.1 with $A_j = \left[u_{11} = u_{21} = u_{(i)} \text{ and } u_{(j)} \text{ is selected in the sample}\right]$, $j = 1, \ldots, x$, then, in samples 1 and 2 we need only choose $n_1 - 1$ and $n_2 - 1$ animals respectively from a total of $R + x - 1$, but n_j animals in samples $j = 3, \ldots, t$ from a total of $R + x$. To establish the probabilities in the form required for Lemma 1.1, we divide this by all possible selections, choosing $n_1 - 1$ animals from $N + R - 1$ for the first sample, *etc*, to get:

$$
\begin{aligned}
&p\left[u_{11}=u_{21}=u_{(i)}\bigcap u_{(2)},\ldots,u_{(x)}\right] \\
&\quad=\frac{\displaystyle\sum_{k=0}^{x-1}(-1)^k\binom{x-1}{k}\binom{R+x-1-k}{n_1-1}\binom{R+x-1-k}{n_2-1}\prod_{j=3}^{t}\binom{R+x-k}{n_j}}{(N+R)\cdot(N+R)\cdot\displaystyle\binom{N+R-1}{n_1-1}\binom{N+R-1}{n_2-1}\prod_{j=3}^{t}\binom{N+R}{n_j}} \\
&\quad=\frac{\displaystyle\sum_{k=0}^{x-1}(-1)^k\frac{x-k}{x}\binom{x}{k}\frac{n_1n_2}{(R+x-k)^2}\binom{R+x-k}{n_1}\binom{R+x-k}{n_2}\prod_{j=3}^{t}\binom{R+x-k}{n_j}}{n_1n_2\displaystyle\binom{N+R}{n_1}\binom{N+R}{n_2}\prod_{j=3}^{t}\binom{N+R}{n_j}} \\
&\quad=\frac{\displaystyle\sum_{k=0}^{x-1}(-1)^k\frac{x-k}{x(R+x-k)^2}\binom{x-1}{k}\prod_{j=1}^{t}\binom{R+x-k}{n_j}}{\displaystyle\prod_{j=1}^{t}\binom{N+R}{n_j}}. \qquad (6.11)
\end{aligned}
$$

Thus, multiplying (6.11) by $x\binom{N}{x}$ gives

$$
p\left[u_{11}=u_{21}\bigcap X=x\right]=\frac{\displaystyle\binom{N}{x}\left[\sum_{k=0}^{x-1}(-1)^k\frac{x-k}{(R+x-k)^2}\binom{x}{k}\prod_{j=1}^{t}\binom{R+x-k}{n_j}\right]}{A(N,\mathbf{n},R)}. \qquad (6.12)
$$

Thus, the generalisation to model M_{fp} of Pathak's (1964) estimate of $\frac{1}{N}$, which he showed to be unbiased with minimum variance under model M_f, is found by substituting (6.8) and (6.12) into (6.10) to get

$$
\hat{N}_P^{-1}(x)=\frac{\displaystyle\sum_{k=0}^{x-1}(-1)^k\frac{x-k}{(R+x-k)^2}\binom{x}{k}\prod_{j=1}^{t}\binom{R+x-k}{n_j}}{x!a(x,\mathbf{n},R)}. \qquad (6.13)
$$

Calculating $x\hat{N}_P^{-1}$ gives an estimate $\hat{C}_P$ of C.

6.4 Goudie estimator

An improved estimator for (6.13) can be achieved under model M_p by the Rao-Blackwell Theorem. Under M_p, the conditional distribution of $\mathbf{n}$ given the sufficient statistics is

$$p(\mathbf{n}|z,x) = \frac{z!a(x,\mathbf{n},R)}{G(z,x,t,Rt)} \tag{6.14}$$

for

$$\mathbf{n} \in \mathbf{n}_{z,x} = \{\mathbf{n}|n_1 + \ldots + n_t = z, n_i \leq x + R, i = 1,\ldots,t\}. \tag{6.15}$$

Thus, if we use the Rao-Blackwell Theorem on (6.13), we get an estimator of $\frac{1}{N}$ that is unbiased, with minimum variance for all $\mathbf{n} \in \mathbf{n}_{z,x}$, under model M_p (Goudie (Personal communication) when $R = 0$). Simplifying (c.f. Feller (1968, p. 58)) gives:

$$\begin{aligned} E[\hat{N}_{-1}|z,x] &= \sum_{\mathbf{n}} \frac{z!}{x!G(z,x,t,Rt)} \sum_{k=0}^{x-1} (-1)^k \frac{x-k}{(R+x-k)^2} \binom{x}{k} \prod_{j=1}^{t} \binom{R+x-k}{n_j} \\ &= \frac{z!}{x!G(z,x,t,Rt)} \sum_{k=0}^{x-1} (-1)^k \frac{x-k}{(R+x-k)^2} \binom{x}{k} \binom{t(R+x-k)}{z}. \end{aligned} \tag{6.16}$$

(Note: This can be done since both summations have a finite number of terms, and so their order can be reversed.)

Thus, on simplifying (6.16), Goudie's estimate of coverage becomes

$$\hat{N}_G^{-1} = \frac{(z-1)!t}{(x-1)!G(z,x,t,Rt)} \sum_{s=1}^{x} \frac{(-1)^{x-s}}{R+s} \binom{x-1}{s-1} \binom{t(R+s)-1}{z-1}, \tag{6.17}$$

and Goudie's coverage estimator is $\hat{C}_G = x\hat{N}_G^{-1}$.

6.5 Computational work

Some computational work has been carried out to check that Pathak's estimator of $\frac{1}{N}$ is in fact unbiased. This was a fairly crude analysis, under Model M_f, choosing some capture histories and calculating expectations. It is evident from Table 6.1 (and others for larger t values) that the claims of unbiasedness hold in the M_f model.

For analysis under model M_{tp}, simulation was carried out using selected beta distributions given in Table 4.1 to generate beta-generated capture probabilities. These

N	**n**=(1,1)		**n**=(2,1)		**n**=(2,2)	
	$\hat{N}_P^{-1}$	$\hat{N}_G^{-1}$	$\hat{N}_P^{-1}$	$\hat{N}_G^{-1}$	$\hat{N}_P^{-1}$	$\hat{N}_G^{-1}$
2	0.50	0.625	0.50	0.50	0.50	0.50
3	0.33	0.50	0.33	0.3611	0.33	0.3611
4	0.25	0.4375	0.25	0.2917	0.25	0.2830
5	0.20	0.40	0.20	0.25	0.20	0.2344
10	0.10	0.325	0.10	0.1667	0.10	0.1343
25	0.04	0.28	0.04	0.1167	0.04	0.0727

Table 6.1: Sample expectations of estimated coverage for the Pathak and Goudie coverage estimators for pre-chosen **n** in cases where t=2.

capture probabilities were used for various values of N and t to assess how each estimator performed under each scenario. The results are given in Tables 6.2 – 6.7 and summarised below.

6.5.1 $N = 50,\ R = 0,\ t = 5$ results

It can be seen that an increase in standard deviation of the capture probabilities increases the sample standard deviation of the true coverage. A doubling of the mean capture probability from 0.05 to 0.1 increases the true coverage by around 80%. Both of these results are roughly as would be expected. The mean estimates from all estimators are below the true mean coverage in all the cases simulated.

The first observation from Tables 6.2 and 6.3 is that Chao et al.'s (1992) second estimator, (6.5), is the best of their three proposed estimators in terms of mean coverage estimate for these trials. It also has the highest sample standard deviation of the three estimators, but this can be explained by the higher average point estimates.

The generalised Pathak coverage estimator has the closest mean estimate to the true coverage mean when $\mu = 0.05$, but for $\mu = 0.1$ $\hat{C}_2$ has the closest mean in two of the three cases. In the first case of Table 6.2 the sample standard deviations of the estimators are as high as three times that of the true sample standard deviation. The generalised Pathak inverse estimator has the highest sample standard deviation of all the estimators in Table 6.2, but this may be a consequence of its higher mean. In Table 6.3, $\hat{C}_2$ has the highest sample standard deviation despite its mean only being the closest to the true mean on only two occasions.

The Goudie estimator has a mean estimate and sample standard deviation that are always below that of the generalised Pathak inverse estimator. In Table 6.2 its mean is above that of $\hat{C}_2$ in all three situations, but below $\hat{C}_2$ in all three situations of Table 6.3. For $\mu = 0.05$ it has a lower sample standard deviation than $\hat{C}_2$ in two of the three cases, despite the higher mean point estimate. In Table 6.3 the Goudie estimator sample standard deviation is always below $\hat{C}_2$, but this may be correlated with the lower mean point estimate in each case. Thus, the Goudie estimator's performance can be considered satisfactory when $N = 50$ and $R = 0$, and be considered especially useful if there is no knowledge of individual capture history or sample sizes.

6.5.2 $N = 50,\ R = 0,\ t = 10$ results

When $N = 50$, increasing the number of samples from 5 to 10 can be seen to increase coverage by around 17.5% when $\mu = 0.05$ and by around 24% when $\mu = 0.1$. Again, all estimators' mean point estimates underestimate the true mean coverage in all cases simulated and have sample standard deviations larger than the true sample standard deviation.

When $t = 10$ there is no clear optimal estimator between $\hat{C}_P$ and $\hat{C}_2$ in terms of mean point estimate. In every case except from the $(0.9, 17.1)$ case, $\hat{C}_P$ has a lower sample standard deviation than $\hat{C}_2$. In the $(0.9, 17.1)$ $\hat{C}_P$ has a mean point estimate that is 3% higher than $\hat{C}_2$, which may explain the former's higher standard deviation in this case.

In every case, $\hat{C}_G$ has a lower mean point estimate than both $\hat{C}_P$ and $\hat{C}_2$. The $\hat{C}_G$ sample standard deviation is always less than or equal to $\hat{C}_P$ and less than $\hat{C}_2$ in the cases simulated. This may be explained by its lower mean point estimate than the other two, however.

Taking both mean and standard deviation into account, the generalised Pathak inverse estimator is preferable to the others simulated when $N = 50$ and $t = 10$.

6.5.3 $N = 50,\ R = 10,\ t = 5$ results

It can be observed from Tables 6.6 – 6.7 that there does not appear to be any systematic difference in the true mean coverage between the M_t and M_{tp} simulations. There is maybe an argument that the true sample standard deviation is slightly smaller when $R = 10$, but the evidence presented here is not conclusive enough.

Table 6.7 offers weak evidence that the mean estimate from $\hat{C}_2$ is slightly improved when $R = 10$. However, this is not supported in Table 6.6. Estimator $\hat{C}_2$ has the closest mean estimate to the true mean out of all those simulated. There is even one case (57.5, 287.85) where its mean overestimates the true mean coverage.

From the results for the cases when $R = 10$, it is evident that $\hat{C}_P$ and $\hat{C}_G$ have negatively biased mean coverage estimations under model M_{tp}. Computation carried out under M_{fp} evaluated that the expected value of the generalised Pathak inverse estimator, $\hat{N}_P^{-1}$, is

$$E\left[\hat{N}_P^{-1}\right] = \frac{1}{N + R\left(2 + \frac{R}{N}\right)}. \tag{6.18}$$

This bias is independent of t and $\mathbf{n}$. However, it is clear that $\hat{C}_P$ should improve asymptotically. The bias should decrease with increasing N and/or decreasing R. This bias also results in $\hat{N}_G^{-1}$ being similarly biased.
Thus, under model M_{tp} the proposed coverage estimator is $\hat{C}_2$.

6.6 Conclusion

Under model M_t, it has been shown that both the generalised Pathak inverse estimator, $\hat{C}_P$, and the Goudie estimator, $\hat{C}_G$ are preferable to the three coverage estimators of Chao et al. (1992) in some of the cases simulated. In particular, $\hat{C}_P$ appears to be the optimal estimator when the capture probability standard deviation is 0.2. This implies that when it is believed that there is strong heterogeneity between the samples, $\hat{C}_P$ should be favoured.
However, under model M_{tp} the generalisation of the Pathak inverse estimator is shown to be biased, with the bias being proportional to $\frac{1}{N}$. Thus, the generalised Pathak inverse estimator, and the Goudie estimator under M_{tp}, are currently of limited use. The estimators of Chao et al. (1992) appear to be unaffected by the inclusion of plants, Thus, plant-capture is not currently recommended for use when estimating sample coverage.

(α, β)	Estimator	Mean estimate	Sample std dev.
(15.15, 287.85)	**True**	**0.2281**	**0.0618**
	Pathak	0.2194	0.1878
	Goudie	0.2172	0.1846
	Chao1	0.1761	0.1532
	Chao2	0.2159	0.1873
	Chao3	0.2133	0.1852
(3.75, 71.25)	**True**	**0.2255**	**0.0766**
	Pathak	0.2080	0.1935
	Goudie	0.1969	0.1799
	Chao1	0.1589	0.1482
	Chao2	0.1955	0.1815
	Chao3	0.1935	0.1769
(0.9, 17.1)	**True**	**0.2284**	**0.1096**
	Pathak	0.2384	0.2538
	Goudie	0.1808	0.1852
	Chao1	0.1401	0.1494
	Chao2	0.1734	0.1847
	Chao3	0.1722	0.1836

Table 6.2: True coverage and mean coverage estimates and sample standard deviations for 1000 realisations of model M_t when $N = 50,\ R = 0$ and $t = 5$, with beta-generated capture probabilities with mean 0.05.

(α, β)	Estimator	Mean estimate	Sample std dev.
(57.5, 517.5)	**True**	**0.4063**	**0.0716**
	Pathak	0.3922	0.1490
	Goudie	0.3906	0.1482
	Chao1	0.3278	0.1330
	Chao2	0.3979	0.1606
	Chao3	0.3902	0.1581
(14.3, 128.7)	**True**	**0.4085**	**0.0786**
	Pathak	0.3913	0.1544
	Goudie	0.3875	0.1525
	Chao1	0.3246	0.1348
	Chao2	0.3934	0.1616
	Chao3	0.3855	0.1583
(3.5, 31.5)	**True**	**0.4119**	**0.1036**
	Pathak	0.3937	0.1721
	Goudie	0.3763	0.1658
	Chao1	0.3164	0.1470
	Chao2	0.3843	0.1763
	Chao3	0.3774	0.1726

Table 6.3: True coverage and mean coverage estimates and sample standard deviations for 1000 realisations of model M_t when $N = 50,\ R = 0$ and $t = 5$, with beta-generated capture probabilities with $\mu = 0.1$.

(α, β)	Estimator	Mean estimate	Sample std dev.
(15.15, 287.85)	**True**	**0.4043**	**0.0733**
	Pathak	0.3884	0.1330
	Goudie	0.3868	0.1325
	Chao1	0.3574	0.1287
	Chao2	0.3903	0.1397
	Chao3	0.3888	0.1391
(3.75, 71.25)	**True**	**0.4018**	**0.0830**
	Pathak	0.3855	0.1470
	Goudie	0.3769	0.1446
	Chao1	0.3469	0.1377
	Chao2	0.3783	0.1486
	Chao3	0.3767	0.1477
(0.9, 17.1)	**True**	**0.4036**	**0.1215**
	Pathak	0.3927	0.1704
	Goudie	0.3598	0.1589
	Chao1	0.3319	0.1512
	Chao2	0.3621	0.1633
	Chao3	0.3606	0.1623

Table 6.4: True coverage and mean coverage estimates and sample standard deviations for 1000 realisations of model M_t when $N = 50,\ R = 0$ and $t = 10$, with beta-generated capture probabilities with mean 0.05.

(α, β)	Estimator	Mean estimate	Sample std dev.
(57.5, 517.5)	**True**	**0.6498**	**0.0699**
	Pathak	0.6453	0.0863
	Goudie	0.6447	0.0861
	Chao1	0.6044	0.0898
	Chao2	0.6467	0.0947
	Chao3	0.6420	0.0940
(14.3, 128.7)	**True**	**0.6534**	**0.0711**
	Pathak	0.6449	0.0920
	Goudie	0.6421	0.0920
	Chao1	0.6034	0.0948
	Chao2	0.6460	0.0994
	Chao3	0.6413	0.0985
(3.5, 31.5)	**True**	**0.6477**	**0.0932**
	Pathak	0.6467	0.1061
	Goudie	0.6348	0.1051
	Chao1	0.5985	0.1078
	Chao2	0.6415	0.1130
	Chao3	0.6370	0.1120

Table 6.5: True coverage and mean coverage estimates and sample standard deviations for 1000 realisations of model M_t when $N = 50,\ R = 0$ and $t = 10$, with beta-generated capture probabilities with mean 0.1.

(α, β)	Estimator	Mean estimate	Sample std dev.
(15.15, 287.85)	**True**	**0.2254**	**0.0628**
	Pathak	0.1504	0.0560
	Goudie	0.1499	0.0558
	Chao1	0.1690	0.1355
	Chao2	0.2078	0.1664
	Chao3	0.2056	0.1649
(3.75, 71.25)	**True**	**0.2231**	**0.0732**
	Pathak	0.1497	0.0621
	Goudie	0.1477	0.0612
	Chao1	0.1584	0.1413
	Chao2	0.1951	0.1742
	Chao3	0.1932	0.1730
(0.9, 17.1)	**True**	**0.2272**	**0.1067**
	Pathak	0.1476	0.0820
	Goudie	0.1410	0.0780
	Chao1	0.1423	0.1452
	Chao2	0.1758	0.1788
	Chao3	0.1745	0.1774

Table 6.6: True coverage and mean coverage estimates and sample standard deviations for 1000 realisations of model M_t when $N = 50,\ R = 10$ and $t = 5$, with beta-generated capture probabilities with mean 0.05.

(α, β)	Estimator	Mean estimate	Sample std dev.
(57.5, 517.5)	**True**	**0.4033**	**0.0695**
	Pathak	0.2773	0.0511
	Goudie	0.2770	0.0511
	Chao1	0.3341	0.1170
	Chao2	0.4045	0.1398
	Chao3	0.3961	0.1388
(14.3, 128.7)	**True**	**0.4103**	**0.0764**
	Pathak	0.2794	0.0561
	Goudie	0.2781	0.0558
	Chao1	0.3292	0.1171
	Chao2	0.3994	0.1410
	Chao3	0.3920	0.1388
(3.5, 31.5)	**True**	**0.4145**	**0.0990**
	Pathak	0.2806	0.0718
	Goudie	0.2754	0.0705
	Chao1	0.3220	0.1306
	Chao2	0.3911	0.1571
	Chao3	0.3840	0.1540

Table 6.7: True coverage and mean coverage estimates and sample standard deviations for 1000 realisations of model M_t when $N = 50,\ R = 10$ and $t = 5$, with beta-generated capture probabilities with mean 0.1.

Chapter 7

General discussion and further work

Chapter 2, working under model M_{fp}, gives the model's probability theory and the generalisation of the estimator of Pathak (1964) to allow for estimation with the inclusion of plants. The chapter also makes use of Berg (1974) to give a variance estimator under model M_{fp}. Also, by a derivation analogous to what Berg (1976, Property 1) gave under M_f, two recurrence relations are given for the calculation of the a-coefficients, (2.4), under M_{fp}, which can be difficult and long to compute. It also compares the approximation given by Pathak (1964, p. 79) to his estimator. An improved approximation may possibly be derived, but the need for such an approximation under M_{fp} can be viewed as no longer crucial. This chapter has shown that the generalised Pathak estimator can be computed for non-unitary sample sizes, reducing the need to approximate. A possible future piece of work would be to derive some special cases of the generalised Pathak estimator, *eg* when $t = 1$ and $R > 0$ or when $t = 2$ and $R \geq 0$. It may be possible to relate these to some standard formulae.

This chapter also gives the calculation of the M_{fp} MLE. A future piece of work could be to derive the asymptotic distribution of this MLE. Goudie et al. (2007) show the asymptotic normality of the model M_p MLE, so it may be that the M_{fp} MLE is also asymptotically normally distributed.

Chapter 2 mainly used trials where the sample sizes were constant between all samples. A future piece of work cold be to extend these results to cases where the sample sizes were different between the samples. This would involve establishing a method of deciding how many captures there should be in each sample. It may be that an ascending or descending number of captures in each sample could be used, or that the number of captures in each sample are determined beforehand via a chosen distribution.

Chapter 3 gives an analysis of the effect that applying the condition that $z > x$ to the closed form estimators has on their estimation. It can be seen that this can cause a lower mean point estimate, which should be taken into account whenever this condition is applied. A lot of theory has been published for models M_0 and M_p, so future work would mainly consist of seeking to generalise these results.
It is shown in Chapter 3 that the generalised Pathak estimator and the M_p CUE have very similar mean estimates and sample standard deviations. Thus, as the generalised Pathak estimator made computation more complex, this could be omitted from future work and one could just compare the M_p CUE with the M_p MLE. This would allow a more complete picture to be presented, which could possibly improve any conclusions made.

Chapter 4 details work carried out under model M_{tp}. The probability theory for this model is given. Also, estimators are computed for this chapter that were derived under simpler models. The first extension that may be possible would be to show that the profile likelihood function under model M_{tp} is unimodal. The proof of the unimodality of the M_p profile likelihood was given in Goudie et al. (2007) and for model M_{fp} in Goudie & Gormley (in submission). It may be that either of these papers can be generalised for such a proof. Also in the paper of Goudie et al. (2007), the asymptotic distribution of the M_p MLE was proven to be normal. It remains to be shown that the same holds under model M_{tp}.
Another point made in this Chapter was the possibility that the normal assumption used for the construction of the M_p and M_{tp} MLE confidence intervals may not hold. In some trials the mean estimated standard deviation for these estimators were larger than their mean point estimate. Thus, one future piece of work could be to use alternative methods of confidence interval estimation, like bootstrapping.

It was shown in Chapter 5 that the M_p CUE had a mean estimate that improved with the inclusion of plants to have an unbiased mean. However, a clear area of future work suggested by this work is the possibility of an improved sparse data estimator for the Chao-type estimator under model M_{tp}. It appears that the information gained from including plants is not being utilised. As such, a fair analysis of the estimator's performance under M_{tp} remains outstanding.

Chapter 6 has shown that the inverse population estimator given by Pathak may be used as a coverage estimator and can be expected to be as good, if not better, than the coverage estimators proposed by Chao et al. (1992). Also given is a new inverse population estimator, $\hat{N}_G^{-1}$. When used to calculate coverage, $\hat{C}_G$ had a mean estimate below that of $\hat{C}_P$ and $\hat{C}_2$ in almost every trial. It was stated in this chapter that plant-capture could not be recommended for estimating coverage, based on the simulations therein. Thus, future work could be to improve the plant-capture theory when estimating coverage, and improve the biased estimators $\hat{C}_P$ and $\hat{C}_G$.

An area that was briefly looked into but not fully covered was a Bayesian approach to plant-capture. Plant-capture lends itself quite naturally to Bayesian methods of estimation, as the practitioner will normally have some prior information about the population under study before the sampling begins, and the information gained from the number of plants captured would assist in deriving posterior distributions. Mark-recapture was first put into a Bayesian framework by Freeman (1972) and Freeman (1973), where he estimated the population size, N, under a sequential recapture framework. Castledine (1981) sought point and interval estimates for N, under models M_0 and M_t using Beta priors for the capture probabilities. Smith (1991) used Bayes, empirical Bayes and Bayes empirical Bayes methods to compute point and interval population size estimates under model M_t. George & Robert (1992) used Gibbs sampling to estimate point estimates of N. Some preliminary work was carried out, aiming to extend George & Robert (1992) to a Bayesian plant-capture scenario under model M_{tp} using Gibbs sampling. This offers a promising area of future research.

Another comment that appears variously throughout this thesis is that the Pathak estimator was possibly considered too difficult to compute in the past. This work has shown that this estimator and its generalisation can now be calculated for the models considered here. However, a future project may be to create a front-end user-friendly interface to the code that I wrote and allow its use by others. This would possibly encourage its use amongst practitioners.

Thus, this thesis has offered some expansion of the plant-capture research work, and has offered some suggestions of further areas of expansion.

Appendix A

Proof of the unimodality of the M_{fp} MLE

To prove that, when $x < z$, the likelihood function is unimodal under model M_{fp}, we first momentarily treat the likelihood function (2.7) as continuous and, without loss of generality, label $n_1 = \max(n_1, \dots, n_t)$. We can then write the score function as

$$\ell'(N; x) = \sum_{i=0}^{x-1} \frac{1}{N-i} - \sum_{i=0}^{n_1-1} \frac{k_{i+1}}{N+R-i} \qquad N \geq x,$$

where k_i is the number of samples at least size i, $(i = 1, \dots, n_1)$. Equivalently

$$\ell'(N; x) = -\sum_{i=0}^{c-1} \frac{k_{i+1}}{N+R-i} - \sum_{i=0}^{n_1-c-1} \frac{k_{i+c+1}-1}{N-i} + \sum_{i=n_1-c}^{x-1} \frac{1}{N-i}, \tag{A.1}$$

where $c = \min\{n_1, R\}$. Note that, by standard mathematical convention, the second sum is zero if $c = n_1$. If there is a stationary point of the likelihood function at N^*, then $\ell'(N^*; x) = 0$. Suppose now that $N' > N^*$. Then $\ell'(N^*; x) - \ell'(N'; x)$ is given by

$$\begin{aligned} & -\sum_{i=0}^{c-1} \frac{k_{i+1}(N'-N^*)}{(N^*+R-i)(N'+R-i)} - \sum_{i=0}^{n_1-c-1} \frac{(k_{i+c+1}-1)(N'-N^*)}{(N^*-i)(N'-i)} + \sum_{i=n_1-c}^{x-1} \frac{N'-N^*}{(N^*-i)(N'-i)} \\ > \; & \frac{-(N'-N^*)}{N'-n_1+c+1} \left\{ \sum_{i=0}^{c-1} \frac{k_{i+1}}{N^*+R-i} + \sum_{i=0}^{n_1-c-1} \frac{k_{i+c+1}-1}{N^*-i} \right\} + \sum_{i=n_1-c}^{x-1} \frac{N'-N^*}{(N^*-i)(N'-i)}. \end{aligned}$$

As N^* gives a root of (A.1), this is equal to

$$\begin{aligned} & \frac{-(N' - N^*)}{N' - n_1 + c + 1} \sum_{i=n_1-c}^{x-1} \frac{1}{N^* - i} + \sum_{i-n_1-c}^{x-1} \frac{N' - N^*}{(N^* - i)(N' - i)} \\ = & \; (N' - N^*) \sum_{i=n_1-c}^{x-1} \frac{i - n_1 + c + 1}{(N^* - i)(N' - i)(N' - n_1 + c + 1)}, \end{aligned}$$

which is positive. Hence, there is no value N' greater than N^* at which the score function takes the value zero, implying the unimodality of the likelihood function. Hence, if non-integer values of the maximum likelihood estimate are permitted, it must be unique. Using different approaches, Pickands & Raghavachari (1987) and Leite, Oishi & de B. Pereira (1988) obtained the latter conclusion for the case where no plants are present.

Bibliography

Amstrup, S. C., McDonald, T. L. & Manly, B. F. J., eds (2005), *Handbook of Capture-Recapture Analysis*, Vol. 1, Princeton University Press.

Arnason, A. N., Kirby, C. W., Schwarz, C. J. & Irvine, J. R. (1996), 'Computer analysis of marking data from stratified populations for estimation of salmonoid escapements and the size of other populations.', *Canadian Technical Report of Fisheries and Aquatic Sciences* **2106**, *vi*+37p.

Ashbridge, J. (1998), Inference for plant-capture, PhD thesis, University of St Andrews.

Ashbridge, J. & Goudie, I. (2009), 'Conditionally unbiased estimation of population size under plant-capture', *Communications in Statistics - Theory and Methods* **38**, 1–12.

Bailey, N. T. J. (1951), 'On estimating the size of mobile populations from recapture data', *Biometrika* **38**, 293–306.

Begon, M. (1979), *Investigating Animal Abundance: capture-recapture for biologists*, Vol. 1, Edward Arnold.

Berg, S. (1974), 'Factorial series distributions, with applications to capture-recapture problems', *Scandanavian Journal of Statistics* **1**, 145–152.

Berg, S. (1976), 'A note on the UMVU estimate in a multiple-recapture census', *Scandanavian Journal of Statistics* **3**, 86–88.

Bishop, Y. M., Fienberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Applications*, Springer.

Borchers, D., Buckland, S. & Zucchini, W. (2002), *Estimating Animal Abundance: Closed Populations*, 1st edn, Springer, London.

Buckland, S., Burnham, K. P. & Augustin, N. H. (1997), 'Model selection: An integral part of inference', *Biometrics* **53**, 603 –618.

Buckland, S., Goudie, I. & Borchers, D. (2000), ‘Wildlife population assessment: Past developments and future directions’, *Biometrics* **56**, 1–20.

Burnham, K. P. & Overton, W. S. (1978), ‘Estimation of the size of a closed population when capture probabilities vary among animals’, *Biometrika* **65**, 625 – 633.

Burnham, K. P., White, G. C. & Anderson, D. R. (1995), ‘Model selection strategy in the analysis of capture-recapture data’, *Biometrics* **51**, 888 – 898.

Casella, G. & George, E. I. (1992), ‘Explaining the Gibbs sampler’, *The American Statistician* **46**, 167–174.

Castledine, B. J. (1981), ‘A Bayesian analysis of multiple-recapture sampling for a closed population’, *Biometrika* **68**, 197–210.

Chao, A. (1989), ‘Estimating population size for sparse data in capture-recapture experiments’, *Biometrics* **45**, 427–438.

Chao, A., Lee, S.-M. & Jeng, S.-L. (1992), ‘Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal’, *Biometrics* **48**, 201–216.

Chapman, D. G. (1951), ‘Some properties of the hypergeometric distribution with applications to zoological sample censuses’, *University of California Publications in Statistics* **1**, 131–160.

Chapman, D. G. (1952), ‘Inverse, multiple and sequential sample censuses’, *Biometrics* **8**, 286–306.

Charalambides, C. A. & Singh, J. (1988), ‘A review of the stirling numbers, their generalizations and statistical applications’, *Communications in Statistics-Theory and Methods* **17**, 2533–2595.

Cormack, R. (1968), ‘The statistics of capture-recapture methods’, *Oceanography and Marine Biology: An Annual Review* **6**, 455–506.

Cormack, R. (1989), ‘Log-linear models for capture-recapture’, *Biometrics* **45**, 395 – 413.

Craig, C. C. (1953), ‘On the utilization of marked specimens in estimating populations of flying insects’, *Biometrika* **40**, 170–176.

Darroch, J. (1958), ‘The multiple-recapture census: I. estimation of a closed population’, *Biometrika* **45**, 343–359.

Darroch, J. & Ratcliff, D. (1980), ‘A note on capture-recapture estimation’, *Biometrics* **36**, 149 – 153.

Dorazio, R. M. & Royle, J. A. (2003), ‘Mixture models for estimating the size of a closed population when capture rates vary among individuals’, *Biometrics* **59**, 351 – 364.

Feller, W. (1968), *An introduction to probability theory and its applications*, Vol. 1, 3rd edn, John Wiley.

Fewster, R. M. & Jupp, P. E. (2009), ‘Inference on population size in binomial detectability models’, *Biometrika* **96**, 805 – 820.

Freeman, P. R. (1972), ‘Sequential estimation of the size of a population’, *Biometrika* **59**, 9–17.

Freeman, P. R. (1973), ‘Sequential recapture’, *Biometrika* **60**, 141–153.

Gazey, W. J. & Staley, M. J. (1986), ‘Population estimation from mark-recapture experiments using a sequential Bayes algorithm’, *Ecology* **67**, 941–951.

George, E. I. & Robert, C. P. (1992), ‘Capture-recapture estimation via Gibbs sampling’, *Biometrika* **79**, 677–683.

Goudie, I. (1995), ‘A plant-capture approach for achieving complete coverage of a population’, *Communications in Statistics - Theory and Methods* **24**, 1293–1305.

Goudie, I. & Ashbridge, J. (2000), ‘A conditionally-unbiased estimator of population size based on plant-capture in continuous time’, *Communications in Statistics - Theory and Methods* **29**, 2605–2619.

Goudie, I. & Ashbridge, J. (2005), ‘A conditionally unbiased estimator for the equal-catchability model’, *Communications in Statistics - Theory and Methods* **34**, 1543–1553.

Goudie, I. & Gormley, R. (in submission), ‘Maximum likelihood estimates for the Schnabel census with plants’.

Goudie, I. & Goudie, M. (2007), ‘Who captures the marks for the Petersen estimator?’, *Journal of the Royal Statistical Society Series A* **170**, 825–839.

Goudie, I., Jupp, P. & Ashbridge, J. (2007), ‘Plant-capture estimation of the size of a homogeneous population’, *Biometrika* **94**, 243–248.

Goudie, I., Pollock, K. H. & Ashbridge, J. (1998), ‘A plant-capture approach for population size estimation in continuous time’, *Communications in Statistics - Theory and Methods* **27**, 433–451.

Gould, H. & Hopper, A. (1962), ‘Operational formulas connected with two generalizations of hermite polynomials’, *Duke Mathematical Journal* **29**, 51–63.

Hammersley, J. (1953), ‘Capture-recapture analysis’, *Biometrika* **40**, 265–278.

Hopper, K., Shinn, M., Laska, E., Meisner, M. & Wanderling, J. (2008), ‘Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods’, *American Journal of Public Health* **98**, 1438–1442.

Huggins, R. M. (1991), ‘On the statistical analysis of capture experiments.’, *Biometrics* **47**, 133 – 140.

Huggins, R. M. (2002), ‘A parametric empirical Bayes approach to the analysis of capture-recapture experiments’, *Australian and New Zealand Journal of Statistics* **44**, 55 – 62.

Hwang, W.-H. & Chao, A. (2002), ‘Continuous-time capture-recapture models with covariates’, *Statistica Sinica* **12**, 1115–1131.

Johnson, N. L., Kotz, S. & Kemp, A. W. (2005), *Univariate Discrete Distributions*, 3rd edn, Wiley, New York.

King, R. & Brooks, S. P. (2001), ‘On the Bayesian analysis of population size’, *Biometrika* **88**, 317 – 336.

King, R. & Brooks, S. P. (2002), ‘Bayesian model discrimination for multiple strata capture-recapture data’, *Biometrika* **89**, 785 – 806.

King, R. & Brooks, S. P. (2008), ‘On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty’, *Biometrics* **64**, 816–824.

King, R., Morgan, B. J. T., Gimenez, O. & Brooks, S. P. (2009), *Bayesian Analysis for Population Ecology*, 1st edn, CRC Press.

Laska, E. M. & Meisner, M. (1993), ‘A plant-capture method for estimating the size of a population from a single sample’, *Biometrics* **49**, 209–220.

Laska, E. M., Meisner, M. & Siegel, C. (1988), ‘Estimating the size of a population from a single sample’, *Biometrics* **44**, 461–472.

Laska, E. M., Meisner, M. & Siegel, C. (1989), 'Correction: Estimating the size of a population from a single sample', *Biometrics* **45**, 1347.

Leite, J., Oishi, J. & de B. Pereira, C. (1988), 'A note on the exact maximum likelihood estimation of the size of a finite and closed population', *Biometrika* **75**, 178 – 180.

Lin, S.-P. & Chao, A. (2005), 'Discrete-time vs. continuous time capture-recapture models', *Journal of the Chinese Statistical Association* **43**, 89–96.

Link, W. A. (2003), 'Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities', *Biometrics* **59**, 1123 – 1130.

Link, W. A. & Barker, R. J. (2009), *Bayesian Inference: with ecological applications*, 1st edn, Academic Press.

Lloyd, C. J. (1994), 'Efficiency of martingale methods in recapture studies', *Biometrika* **81**, 305 – 315.

Martin, E., Laska, E. M., Hopper, K., Meisner, M. & Wanderling, J. (1997), 'Issues in the use of a plant-capture method for estimating the size of the street dwelling population', *Journal of Official Statistics* **13**, 59–73.

McCarthy, M. A. (2007), *Bayesian Methods for Ecology*, 1st edn, Cambridge University Press.

Otis, D. L., Burnham, K. P., White, G. C. & Anderson, D. R. (1978), 'Statistical inference from capture data on closed animal populations', *Wildlife Monographs: A Publication of The Wildlife Society* **62**, 135.

Pathak, P. K. (1961), 'On the Evaluation of Moments of Distinct Units in a Sample.', *Sankhya, The Indian Statistical Journal* **23**, 415–420.

Pathak, P. K. (1964), 'On estimating the size of a population and its inverse by capture mark method', *Sankhya, The Indian Statistical Journal* **A26**, 75–80.

Pickands, J. & Raghavachari, M. (1987), 'Exact and asymptotic inference for the size of a population', *Biometrics* **74**, 355–363.

Pledger, S. (2000), 'Unified maximum likelihood estimates for closed capture-recapture models using mixtures', *Biometrics* **56**, 434–442.

Pollock, K. H. (2000), 'Capture-recapture models', *Journal of the American Statistical Association* **95**, 293–296.

Rupp, R. S. (1966), 'Generalized equation for the ratio method of estimating population abundance', *The Journal of Wildlife Management* **30**, 523–526.

Schnabel, Z. E. (1938), 'The estimation of total fish population of a lake', *The American Mathematical Monthly* **45**, 348–352.

Schwarz, C. J. & Seber, G. A. F. (1999), 'Estimating animal abundance: Review III', *Statistical Science* **14**, 427–456.

Seber, G. A. F. (1982), *The estimation of animal abundance and related parameters*, 2nd edn, Charles Griffen and Co.

Smith, P. J. (1991), 'Bayesian analyses for a multiple capture-recapture model', *Biometrika* **78**, 399–407.

Stanley, T. R. & Burnham, K. P. (1999), 'A goodness-of-fit test for capture-recapture model mt under closure', *Biometrics* **55**, 366–375.
URL: *http://www.jstor.org/stable/2533781*

Stanley, T. R. & Richard, J. D. (2005), 'A program for testing capture-recapture data for closure', *Wildlife Society Bulletin* **33**, 782 – 785.

White, G. C., Anderson, D. R., Burnham, K. P. & Otis, D. L. (1982), *Capture-Recapture and Removal Methods for Sampling Closed Populations*, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

Williams, B. K., Nichols, J. D. & Conroy, M. J. (2002), *Analysis and Management of Animal Populations*, 1st edn, Academic Press, San Diego, CA.

Wilson, R. M. & Collins, M. F. (1992), 'Capture-recapture estimation with samples of size one using frequency data', *Biometrika* **79**, 543 – 553.

Yip, P. S. F. (1996), 'Effect of plant-capture in a capture-recapture experiment', *Communications in Statistics - Theory and Methods* **25**, 2025–2038.

Zelterman, D. (1988), 'Robust estimation in truncated discrete distributions with application to capture- recapture experiments.', *Journal of Statistical Planning and Inference* **18**, 225 –237.