https://doi.org/10.26219/heal.aueb.5897

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

CAPTURE RE-CAPTURE TECHNIQUES FOR ESTIMATING THE SPECIES RICHNESS

By

Alexander {Oleg}. Tsaloukidis

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

June 2018

https://doi.org/10.26219/heal.aueb.5897

https://doi.org/10.26219/heal.aueb.5897

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

CAPTURE RE-CAPTURE TECHNIQUES FOR ESTIMATING THE SPECIES RICHNESS

Αλέξανδρος Όλεγκ Τσαλουκίδης

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούνιος 2018

https://doi.org/10.26219/heal.aueb.5897

https://doi.org/10.26219/heal.aueb.5897

ABSTRACT

Alexander Tsaloukidis

CAPTURE RE-CAPTURE TECHNIQUES FOR ESTIMATING THE SPECIES RICHNESS

June 2018

The number of species in a biological community, usually referred to as species richness, represents the simplest and most direct indicator of ecological diversity. Since scientists cannot detect each plant or animal in a region, species richness is an unknown parameter of the community under study. Chiarucci et al (2018) proposed a new method, called the K_{ED} estimator, to improve the sample-based estimation of species richness and adopt it to estimate species richness of vascular plants in the Maremma Regional Park, Italy. This project proposes a new method for estimating species richness based on the stochastic approach and models. The new approach provides flexibility in the range of models that can be fitted and a likelihood framework for model selection and estimation of uncertainty. For the Maremma Park case study, the new method provides estimates greater than K_{ED} which is close to the number of species detected.

https://doi.org/10.26219/heal.aueb.5897

https://doi.org/10.26219/heal.aueb.5897

TABLE OF CONTENTS

CHAPTER	Page
ABSTRACT	i
LIST OF TABLES	v
LIST OF FIGURES	vii
1. INTRODUCTION	1
2. THE K_{ED} ESTIMATOR	2
3. STOCHASTIC APPROACH	4
4. THE IMPLEMENTATION OF THE STOCHASTIC APPROACH	13
4.1. Descriptive analysis	13
4.2. Estimation using M_0 model	15
4.3. Estimation using $M_{2\text{-}b}$ model	18
4.4. Estimation using $M_{3\text{-}b}$ model	20
5. BOOTSTRAP IMPLIMENTATION	24
5.1. The Bootstrap algorithm	24
5.2. Bootstrap for model M_0	26
5.3. Bootstrap for model $M_{2\text{-}b}$	27
5.4. Bootstrap for model $M_{3\text{-}b}$	28
6. MONTE-CARLO IMPLIMENTATION	29
6.1. The Monte-Carlo algorithm	29
6.2. Monte-Carlo for model M_0	29
6.3. Monte-Carlo for model $M_{2\text{-}b}$	30
6.4. Monte-Carlo for model $M_{3\text{-}b}$	31
7. MODELS WITH G>3	33
8. MODEL AVERAGING	36
9. THE M_t MODEL	38
10. THE BETA-BINOMIAL MODEL	40
11. CONCLUSION/FURTHER RESEARCH	42

https://doi.org/10.26219/heal.aueb.5897

https://doi.org/10.26219/heal.aueb.5897

LIST OF TABLES

Table	Page
2.1: *The structure of the available data*	2
3.1: *The essential quantities for the Schnabel experiment involving n samples and r individuals, where I(A) is the indicator function, equal to unity if A is true*	7
4.1: *Descriptive measurements for the number of times the species were observed during the study*	13
4.2: *The frequencies f_i of the number of species observed i times*	13
4.3: *Descriptive measurements for the observed number of species s_i per sample*	14
4.4: *The results derived from the M_0*	17
4.5: *The results derived from model $M_{2\text{-}b}$*	19
4.6: *Maximum likelihood estimates and estimates of uncertainty for the parameters p_1, p_2 and w*	19
4.7: *The results derived from model $M_{3\text{-}b}$*	22
4.8: *Maximum likelihood estimates and estimates of uncertainty for the parameters p_1, p_2, p_3, w_1 and w_2*	22
5.1: *Summary results from the implementation of the bootstrap procedure for the M_0 model*	26
5.2: *Summary results from the implementation of the bootstrap procedure for the $M_{2\text{-}b}$ model*	27
5.3: *Summary results from the implementation of the bootstrap procedure for the $M_{3\text{-}b}$ model*	28
6.1: *Summary results from the implementation of the Monte-Carlo procedure for the M_0 model*	29
6.2: *Summary results from the implementation of the Monte-Carlo procedure for the $M_{2\text{-}b}$ model*	31
6.3: *Summary results from the implementation of the Monte-Carlo procedure for the $M_{3\text{-}b}$ model*	32
7.1: *The estimation of N for each of the models*	33
7.2: *The estimation of the parameters p_g and w_g and their standard errors for each model*	34
7.3: *Estimation of N from model $M_{7\text{-}b}$*	34
8.1: *The AIC weights for each model*	36
8.2: *The estimation of N as well as the standard error and a 95% c.i. based on the lowest AIC and BIC values*	37
9.1: *The estimation of N, the standard error and a 95% c.i. from the M_t model*	39

https://doi.org/10.26219/heal.aueb.5897

https://doi.org/10.26219/heal.aueb.5897

LIST OF FIGURES

Figure	Page
4.1: *The estimated distribution for the x_j observed values*	14
4.2: *The estimated distribution for the s_i observed values*	15
4.3: *Profile log-likelihood function for N from model M_0*	17
4.4: *Construction of 95% profile log likelihood confidence interval for N from model M_0*	17
4.5: *Profile log Likelihood function for N from model $M_{2\text{-}b}$*	19
4.6: *Construction of 95% profile log likelihood confidence interval for N from model $M_{2\text{-}b}$*	20
4.7: *Profile log-likelihood function for N from model $M_{3\text{-}b}$*	22
4.8: *Construction of 95% profile log likelihood confidence interval for N from model $M_{3\text{-}b}$*	23
5.1: *Histogram of the $\widehat{N}_{ML,M_0}$ from 10000 bootstrap replications together with a normal distribution based on the mean and standard deviation of the bootstrap estimates. The value $\widehat{N}_{ML,M_0}$=979.14 is also shown*	26
5.2: *Histogram of the $\widehat{N}_{ML,M_{2-b}}$ from 1000 bootstrap replications together with a normal distribution based on the mean and standard deviation of the bootstrap estimates. The value $\widehat{N}_{ML,M_{2-b}}$=1147.87 is also shown*	27
5.3: *Histogram of the $\widehat{N}_{ML,M_{3-b}}$ from 1000 bootstrap replications together with a normal distribution based on the mean and standard deviation of the bootstrap estimates. The value $\widehat{N}_{ML,M_{3-b}}$=1282 is also shown*	28
6.1: *Histogram of the $\widehat{N}_{ML,M_0}$ from 10000 Monte-Carlo replications together with a normal distribution based on the mean and standard deviation of the Monte-Carlo estimates. The value $\widehat{N}_{ML,M_0}$=979.14 is also shown* **30**	
6.2: *Histogram of the $\widehat{N}_{ML,M_{2-b}}$ from 1000 Monte-Carlo replications together with a normal distribution based on the mean and standard deviation of the Monte-Carlo estimates. The value $\widehat{N}_{ML,M_{2-b}}$=1147.87 is also shown*	31
6.3: *Histogram of the $\widehat{N}_{ML,M_{3-b}}$ from 1000 Monte-Carlo replications together with a normal distribution based on the mean and standard deviation of the Monte-Carlo estimates. The value $\widehat{N}_{ML,M_{3-b}}$=1282 is also shown*	32
7.1: *BIC and the AIC values for models ($M_{1\text{-}b}$=M_0)*	35

https://doi.org/10.26219/heal.aueb.5897

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 1.

INTRODUCTION

Consider a plant community within a delineated study area. From a statistical point of view the community constitutes a without frame population of N units spread over the area. Under probabilistic sampling, the estimation of species richness can be performed by using either abundance data or presence-absence data (Chiarucci et al (2018)). Capture re-capture is a method commonly used in ecology to estimate an animal population's size (or species richness). A portion of the population is captured, marked, and released. Later, another portion is captured and the number of marked individuals (or different observed species) within the sample is counted. Since the number of marked individuals within the second sample should be proportional to the number of marked individuals in the whole population, an estimate of the total population size can be obtained by dividing the number of marked individuals by the proportion of marked individuals in the second sample. The method is most useful when it is not practical to count all the individuals in the population. Other names for this method, or closely related methods, include capture-recapture, capture-mark-recapture, mark-recapture, sight-resight, mark-release-recapture, multiple systems estimation, band recovery, the Petersen method, and the Lincoln method.

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 2.

THE K_{ED} ESTIMATOR

If K species are present in the community, each assemblage of plants of the same species may be viewed as unit, in such a way that the complete species list can be viewed as a population. Ideally, referring to species by their identifying numerical labels, the complete list of species can be represented by the set **C(U)**=$\{1,\ldots,K\}$, henceforth denoted for brevity by **C**, while K represents the species richness. The sampling scheme adopted to select plants, univocally determines the first-order inclusion probabilities of species $\theta_1,\ldots,\theta_K$. Because the study area cannot be adequately sampled by means of only one plot or transect, n independent replications of the sampling scheme (Barabesi and Fattorini, 1998) are usually performed, giving rise to n samples of plants $S_1,\ldots,S_n$ which in turn, give rise to n samples of species $Q_1,\ldots,Q_n$. Owing to the independence of the replications the probability that he species j enters the pooled sample $Q_{(n)}$ (i.e. it is detected during the whole sample study) turns out to be $\tau_j = 1-(1-\theta_j)^n$. (Chiarucci et al, 2018)

The structure of the available data is shown in the Table 2.1, and it is expressed as:

$$I_{j,i} = \begin{cases} 1, & if\ a\ unit\ of\ the\ j^{th} species\ was\ observed\ during\ the\ i^{th}\ sample \\ 0, & otherwise \end{cases}$$

	Sample (List)				Floristic List
Species	S_1	. . .	S_{n-1}	S_n	$S_{floristic}$
1	$I_{1,1}$	. . .	$I_{1,n-1}$	$I_{1,n}$	$I_{1,\ floristic}$
2	$I_{2,1}$	. . .	$I_{2,n-1}$	$I_{2,n}$	$I_{2,\ floristic}$
.	.	. . .	.	.	.
.	.	. . .	.	.	.
.	.	. . .	.	.	.
r-1	$I_{r-1,1}$	. . .	$I_{r-1,n-1}$	$I_{r-1,n}$	$I_{r-1,\ floristic}$
r	$I_{r,1}$	. . .	$I_{r,n-1}$	$I_{r,n}$	$I_{r,\ floristic}$

Table 2.1: *The structure of the available data*

https://doi.org/10.26219/heal.aueb.5897

Assuming that there are *n* lists and one floristic list, the K_{ED} estimator proposed by Chiarucci et al (2018), is based on computing the τ_j probabilities using the *n* lists. The measurements $1/\tau_j$ can be regarded as the weight of the j^{th} species. So in order to compute the K_{ED} estimator, the weights of the j^{th} species, which wasn't observed in the floristic list, are added to the number of the species observed in the floristic list.

The τ_j is the probability that species *j* enters the sample $Q_{(n)}$ (i.e. it is detected during the whole sample survey), given that the replications are independent and it is given by the formula:

$$\tau_j = 1 - (1 - \theta_j)^n$$

Where θ_js are given by the formula:

$$\theta_j = \frac{(\sum_{i=1}^{i=n} I_{j,i}) + 1}{n+1},$$

The unit is added, due to the fact that τ_j probabilities should be strictly positive since for the computation of the K_{ED} estimator the measurement $1/\tau_j$ is used.

The final formula for the K_{ED} estimator can be expressed as:

$$\widehat{K}_{ED} = r_{floristic} + \sum_{j=1}^{r} \left[\left(1 - I_{j,floristic}\right) * \frac{1}{\tau_j} \right]$$

Where $r_{floristic} = \sum_{j=1}^{r} I_{j,floristic}$

The K_{ED} estimator is biased with expectation and variance which cannot be expressed in closed forms. Also as it can be noticed from the table that this estimator doesn't offer much of information, since its estimation doesn't differ a lot from the number of different species observed during the study; therefore the stochastic approach will be proposed in order to estimate the total number of different species in the area of study.

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 3.

STOCHASTIC APPROACH

In a typical capture-recapture experiment in ecology, we place traps or nets in the study area and sample the population several times. At the first trapping sample a number of animals are captured; the animals are uniquely tagged or marked and released into the population. For the Maremma Regional Park study, the species can be considered as "animals", the process of marking is the identifying to which species each of the observed plants belongs and the sample as the plot; the aim is to estimate how many different species exist in the whole area which is technically the population as mentioned above.

Capture-recapture models are generally classified as either closed population or open populations models. In a closed population, the size of the population N, which is the main parameter of interest, is assumed to be constant over the trapping/observation times. Open population may have demographic changes or migration. For the given problem, it would be more appropriate to assume closed population models, since it is reasonable to expect that the variety of species in the area has not changed drastically during the sampling period.

The simplest mark-recapture experiment, known as the Lincoln-Petersen procedure, consists of only two samples and provides the most basic estimator for estimating the size N. A sample of size s_1 is captured, marked and released back into the population or equivalently s_1 can be considered the number of different species observed in a plot/sample. The marked proportion of the population is therefore:

$$\frac{s_1}{N}$$

The observation of second of a plot/sample results in s_2 different species from which m_2 species were observed in the s_1 sample. Assuming that the observed marked proportion in the sample is equal to the observed marked proportion in the population,

https://doi.org/10.26219/heal.aueb.5897

$$\frac{m_2}{s_2} \approx \frac{s_1}{N},$$

yields the Lincoln-Petersen estimator for population size/variety of species

$$\widehat{N} = \frac{s_1 s_2}{m_2}.$$

The estimator is intuitive and can be considered as the second sample size, s_2 divided by the capture probability on the second occasion ($\hat{p}_2 = \frac{m_2}{s_1}$).

Furthermore, it is the maximum-likelihood estimator under a number of probabilistic models, including the multinomial and the hypergeometric (Karandinos, 2007). However, the estimator may have an infinite expectation for small samples if m_2 has a finite probability of being zero. Chapman (1951) showed that bias is reduced by setting:

$$N^* = \frac{(s_1+1)(s_2+1)}{m_2+1} - 1.$$

Confidence intervals for Lincoln-Petersen estimates can be constructed in various ways, including asymptotic normality of N^* (Seber, 1982), assuming normality for log N^* (Skalski and Robson, 1992), a method based on the estimated number of species not observed (Chao, 1987) and the use of profile likelihoods. Charts are available for choosing s_1 and s_2 to achieve prescribed accuracy of $\widehat{N}$ (see Seber, 1973, Chap. 3).

The assumptions of the Lincoln-Petersen estimator, and common to all closed population models, are:

1. the population is closed
2. all animals/species are equally likely to be captured/observed in a given sample
3. marks are not lost or overlooked.

An approach in which randomness is related to how the species are distributed in the area considers the N species in the population as N independent multinomial trials each with the same probability of belonging to a given capture-recapture category. In the Petersen experiment there are four

https://doi.org/10.26219/heal.aueb.5897

categories: observed in sample 1 only, observed in sample 2 only, observed in both samples and observed in neither sample, with respective probabilities P_{10}, P_{01}, P_{11} and Q (=1- P_{10}- P_{01}- P_{11}=1-P). Thus, if α_{10} (=s_1 - m_2), α_{01} (=s_2 - m_2), α_{11} (= m_2), and N- α_{10}- -α_{01}- α_{11} (=N-r, where r is the total number of different species observed) are the numbers observed in the four categories, then the joint probability function of the random variables α_{10}, α_{01}, and α_{11} is

$$f(\alpha_{10},\alpha_{01},\alpha_{11}) = \frac{N!}{\alpha_{10}!\alpha_{01}!\alpha_{11}!(N-r)!} \times P_{10}^{\alpha_{10}} P_{01}^{\alpha_{01}} P_{11}^{\alpha_{11}} Q^{N-r}.$$

If p_i (=1-q_i) is the probability of capture/observe in sample i, and samples are independent so that capture/observe in sample 1 does not affect capture/observe in sample 2, then $P_{10} = p_1 q_2$, $P_{01} = q_1 p_2$, $P_{11} = p_1 p_2$ and $Q = q_1 q_2$. Now the triple (α_{10}, α_{01}, α_{11}) determines the triple (α_{11}, α_{10}+ α_{11}, α_{01}+ α_{11})≡(m_2, s_1, s_2,) and vice versa, so that substituting for the P's gives us

$$f(\alpha_{10},\alpha_{01},\alpha_{11}) = f(m_2,s_1,s_2) = \frac{N!}{(s_1-m_2)!(s_2-m_2)!m_2!(N-r)!} p_1^{s_1} q_1^{N-s_1} p_2^{n_2} q_2^{N-s_2} \quad (1)$$

$$= f(m_2 \mid s_1,s_2) \prod_{i=1}^{2} \left\{ \binom{N}{s_i} p_i^{s_i} q_i^{N-s_i} \right\}$$

$$= f(m_2 \mid s_1,s_2) f(s_1,s_2)$$

where $f(m_2 \mid s_1,s_2) = \binom{s_1}{m_2}\binom{N-s_1}{s_2-m_2} / \binom{N}{s_2}$ is the hypergeometric probability function

THE LIKELIHOOD FUNCTION. Formally, the likelihood function is the joint probability density function of the sample data. In the context here, it is a function of the integer-valued parameter N and the real-valued parameter p (the vector containing all the probability parameters necessary to the model), given the discrete sample data contained in the X matrix (a matrix of zeros and ones as defined in Table 1). The notation is not as complex as it may seem. For example, $L(N,p|X)$ denotes the likelihood function of the unknown parameters N and p, given a specific set of sample data contained in the X matrix. The likelihood function is a formal way to express quantitatively the relative "likeliness" of several values that may be considered as candidates for $\widehat{N}$. The maximum likelihood method selects as the value for $\widehat{N}$ the most

https://doi.org/10.26219/heal.aueb.5897

likely one, on the basis of the available data (hence the name, maximum likelihood). The use of likelihood theory and the maximum likelihood method extends our intuition and ability to make inductive inference. The likelihood function may be difficult to deal with analytically because it involves products of often complicated terms.

The Schnabel census

A natural extension for the Peterson experiment is the so-called Schnabel experiment or multiple recapture censuses in which n (n >2) consecutive samples are taken from the population. If s_i different species are observed in sample/plot S_i, and m_i of these are found to have been observed on a previous sampling occasion, then u_i (=s_i - m_i) unobserved species in previous samples/plots are now observed. Table 3.1 shows the basic quantities for this experiment,

i	s_i	f_i
1	$\sum_{j=1}^{r} I_{j,1}$	$\sum_{j=1}^{r} I(x_j = 1)$
2	$\sum_{j=1}^{r} I_{j,2}$	$\sum_{j=1}^{r} I(x_j = 2)$
.	.	.
.	.	.
.	.	.
n	$\sum_{j=1}^{r} I_{j,n}$	$\sum_{j=1}^{r} I(x_j = n)$

Table 3.1: *The essential quantities for the Schnabel experiment involving n samples and r individuals, where I(A) is the indicator function, equal to unity if A is true*

where $x_j = \sum_{i=1}^{n} I_{j,i}$ is the number of times species j was observed during the study, therefore the measurement f_i is the number of species which were observed i times, $i=1,\ldots,n$.

Now the joint probability (1) can be generalized as:

$$\Pr(\{\alpha_\omega\}) = \frac{N!}{\prod_\omega \alpha_\omega!(N-r)!} Q^{N-r} \prod_\omega P_\omega^{\alpha_\omega} = \frac{N!}{\prod_\omega \alpha_\omega!(N-r)!} \prod_{i=1}^{n} p_i^{s_i} q_i^{N-s_i} \qquad (2)$$

https://doi.org/10.26219/heal.aueb.5897

where N is the total number of different species, r is the number of different species observed in the samples/plots, and α_ω denotes the frequency for observable capture history ω.

The assumptions underlying the Petersen method must apply to all the samples in a Schnabel census so that any departures from these assumptions can seriously affect the validity of $\widehat{N}$. When a Schnabel census is used to estimate animals' populations, variation in catchability seems to be a fact of life. Otis et al (1978) Devised a basis of models for estimating N, where these allow capture probabilities to vary with respect to one or more of the factors of time, behavior response, and individual response. In particular, they proposed the following eight models:

- $\boldsymbol{M_0}$: no variation
- $\boldsymbol{M_t}$: variation with trapping occasion or time
- $\boldsymbol{M_b}$: variation by behavioral responses, e.g. trap shyness
- $\boldsymbol{M_h}$: variation by individual response or heterogeneity
- Various combinations of the above models: $\boldsymbol{M_{tb}}$, $\boldsymbol{M_{bh}}$, $\boldsymbol{M_{th}}$, $\boldsymbol{M_{tbh}}$

Since the focus of the research is to estimate the number of different plant species in the Maremma Regional Park, the model which assumes variation by behavioral responses and all of its combinations with the other models, has no relevance for obvious reasons.

If p_{ji} is the probability the j^{th} species ($j = 1,2,\ldots,N$) is observed in the i^{th} sample ($i = 1,2,\ldots,n$), and we can assume that the species are independent of one another as far as observing is concerned, then the likelihood function is:

$$\prod_{j=1}^{N}\prod_{i=1}^{n} p_{ji}^{I_{ji}} (1-p_{ji})^{1-I_{ji}}$$

The various models can now be described mathematically by specifying p_{ji}.

Model M_0: $p_{ij} = p$

https://doi.org/10.26219/heal.aueb.5897

This is the simplest model where the capture probability is constant over capture/observe probability occasions. It allows for no sources of variation in its capture probabilities.

Model M_0 is valuable primarily as a necessary starting point for testing assumptions about capture/observing probabilities. We also use it to introduce many ideas about analysis, such as confidence intervals, that will apply to all estimators. In terms of a restriction on the most general possible model, Model M_0 is equivalent to the assumption that $p_{ji} = p$ for every specie at risk of being observed ($j = 1,...,N$) on every plot ($i = 1,...,n$). There are only two parameters in the model, (N, p), and the joint probability distribution (likelihood) for the data can be written as:

$$\Pr(\{x_\omega\} \mid N, p) = \frac{N!}{\prod_\omega \alpha_\omega!(N-r)!} p^{s\cdot}(1-p)^{nN-s\cdot} \quad (3)$$

where $s\cdot = \sum_{i=1}^{n} s_i$.

The maximization of the likelihood above with respect to N and p will provide maximum likelihood estimates for the number of species and capture/observe probability, which according to this model is constant; this optimization can be achieved using numerical methods such us “BFGS”. A large-sample variance for $\widehat{N}$ is given by (Darroch 1958)

$$Var(\widehat{N}) = \frac{N}{(1-p)^{-n} + (n-1) - n(1-p)^{-1}}$$

Model M_t: $p_{ji} = p_i$

This model allows the capture probabilities to vary with time and reduces to the Lincoln-Petersen model when $n=2$. In general, there are $n+1$ parameters, (N, $p_1,\ldots,p_n$), and the (multinomial) likelihood is given by Equation (2). Under Model M_t we expect a significant variation in the numbers of captures at each sampling occasion. An iterative procedure is needed to find the maximum likelihood estimates of the parameters; the maximum likelihood estimate, $\widehat{N}$,

https://doi.org/10.26219/heal.aueb.5897

is close to the unique root, greater than *r*, of the $(n\text{-}1)^{th}$-degree polynomial (Seber, 1980)

$$(1 - \frac{r}{N}) = \prod_{i=1}^{n}(1 - \frac{s_i}{N}).$$

The left side of this equation estimates the probability that an animal/species is not observed during the study and the right side is the product of estimates of not being observed on each sampling occasion. The variance of $\widehat{N}$ is approximately given by:

$$Var(\widehat{N}) = \frac{N}{\prod_{i=1}^{n}(1 - p_i)^{-n} + (n - 1) - \sum_{i=1}^{n}(1 - p_i)^{-1}}$$

Model M_h: $p_{ji} = p_j$

Capture/Observe probabilities often vary by animal, sometimes for obvious reasons (differences in species, sex, or age), but there also may be unrecognized sources of variation in capture probability by animal (social dominance, number and placement of traps in an animal's home range, or innate level of activity). Because these factors result in capture probabilities that vary among animals, we refer to this source of variation as heterogeneity. Numerous studies reported in the ecological literature clearly show heterogeneity and other sources of variation in capture probabilities for a wide range of species and many types of studies. Examples include Young et al. (1952), Tanaka (1956), Crowcroft and Jeffers (1961), Huber (1962), Edwards and Eberhardt (1967), Bailey (1969), Gliwicz (1970), Carothers (1973a), Beukema and de Vos (1974), Jensen (1975), and Montgomery (1979). In studies where the true population size was known, the commonly used estimators were biased severely by heterogeneity of capture probabilities: the estimates were very much too low. (See, for example, Edwards and Eberhardt(1967) and Carothers(1973a)).

Computer simulation studies have also shown that heterogeneity can cause substantial negative bias in the commonly used estimators. (See, for example, Burnham and Overton, 1969; Manly, 1970; Gilbert, 1973; Carothers, 1973b

https://doi.org/10.26219/heal.aueb.5897

and Otis et al., 1978). In spite of the evidence that heterogeneity exists and invalidates the usual estimates under Model M_t, only recently has Model M_h been formalized and has an estimator been derived for it, because it is a difficult model. (See, Burnham, 1972; Otis et al., 1978:33-37; and Burnham and Overton, 1978).

Model M_h assumes that the j^{th} species has its own unique observing probability, p_j, which is constant over all sampling occasions. The model is thus parameterized in terms of N+1 parameters, (N, p_1,…,p_N), but the likelihood is difficult to handle without any restrictions on the p. One plausible approach is to assume a parametric distribution *f(p)*, which can be either continuous or discrete, for the observing probabilities so that the *p* are a random sample from this distribution. The corresponding statistical model can be described in terms of the number f_i of species being observed exactly i times in the study:

$$\Pr(f_1, \dots, f_n|F) = \frac{N!}{\prod_i f_i!(N-r)!}\pi_0^{N-r}\prod_{i=1}^{n}\pi_i^{f_i} \qquad (4)$$

where π_i is the expected probability that a species is observed exactly *i* times. If the parametric distribution *f(p)* is continuous, π_i is given as:

$$\pi_i = \int_0^1 \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i}dF(p),\ i=0,\dots,n,$$

while if the parametric distribution *f(p)* is discrete, π_i is given as:

$$\pi_i = \sum_{g=1}^{G} w_g\left(\frac{n!}{(n-i)!i!}p^i(1-p)^{n-i}\right),\ i=0,\dots,n.$$

The latter expression is mixture of *G* binomial distributions. In this paper the cases for *G=1,..,7* will be considered.

If the distribution *f(p)* is point mass, model M_h becomes identical to model M_0, therefore in order to make the likelihood for model M_0 comparable with the likelihoods of the M_h models which use the mixture of distributions, model M_0 will be fitted via Eq. 4 than Eq. 3.

https://doi.org/10.26219/heal.aueb.5897

PROFILE LIKELIHOOD: The idea is to invert a likelihood-ratio test to obtain a CI for the parameters of a model. Consider a statistical model with parameters θ and δ where θ is a scalar parameter of interest and δ is the (vector of) additional parameter(s) in the model. Denote by $L(\theta,\delta)$ the likelihood function, and by (θ^*,δ^*) the maximum likelihood (ML) estimates. The likelihood ratio test statistic (G^2) of the hypothesis H_0: $\theta=\theta_0$ (where θ_0 is a fixed value) equals the drop in $2lnL$ between the "full" model and the reduced model with θ fixed at θ_0, i.e. $G^2=2(lnL(\theta^*,\delta^*)-lnL(\theta_0,\delta_0^*))$, where δ_0^* is the maximum likelihood estimate of the reduced model. Alternatively, we may express the test statistic in terms of the profile likelihood function L_1 for the parameter which is obtained from the usual likelihood function by maximizing over the parameter δ, i.e. $L_1(\delta)=max_\delta L(\theta,\delta)$. Then we *have* $G^2=2(lnL_1(\theta^*)-lnL_1(\theta_0))$. A 95% CI for θ consists of those values of θ_0 for which the test is non-significant at significance level 0.05; this is the case when G^2 does not exceed 3.84 (95%-percentile of the $\chi^2(1)$ distribution). Thus, the CI consists of the θ_0-values for which $lnL_1(\theta_0) \geq lnL_1(\theta^*)$ $3.84/2=lnL(\theta^*,\delta^*)-1.92$. (H. Stryhn, J. Christensen). The idea generalizes to vector θ.

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 4.

THE IMPLEMENTATION OF THE STOCHASTIC APPROACH

4.1. Descriptive analysis

The Maremma Park study involves 90 species lists from probabilistic sampling and one list from purposive sampling. In total, 962 species were observed. We do not distinguish between lists in this work. From the available data the average value of $x_j = \sum_{i=1}^{n} I_{j,i}$, that is the average number of times the species were observed during the study, is 3.96, the median is 2, the standard deviation is 5.39 and r is 962; also $\widehat{K}_{ED} = 973.5$. Furthermore the maximum number of times a species is observed is 53 (Table 4.1). Table 4.2 shows the frequency f_i of the number of species observed i times. In addition, Figure 4.1 shows the estimated distribution for the observed x_j values.

Average	Median	Standard Deviation	Maximum	r	$\widehat{K}_{ED}$
3.96	2	5.39	53	962	973.5

Table 4.1: *Descriptive measurements for the number of times the species were observed during the study*

i	f_i	i	f_i	i	f_i	i	f_i	i	f_i	i	f_i
1	419	7	21	13	8	19	1	25	2	37	1
2	153	8	23	14	9	20	2	26	1	46	1
3	102	9	18	15	7	21	4	30	1	53	1
4	49	10	17	16	4	22	2	31	2		
5	41	11	14	17	1	23	5	32	2		
6	35	12	8	18	5	24	2	35	1		

Table 4.2: *The frequencies f_i of the number of species observed i times*

https://doi.org/10.26219/heal.aueb.5897

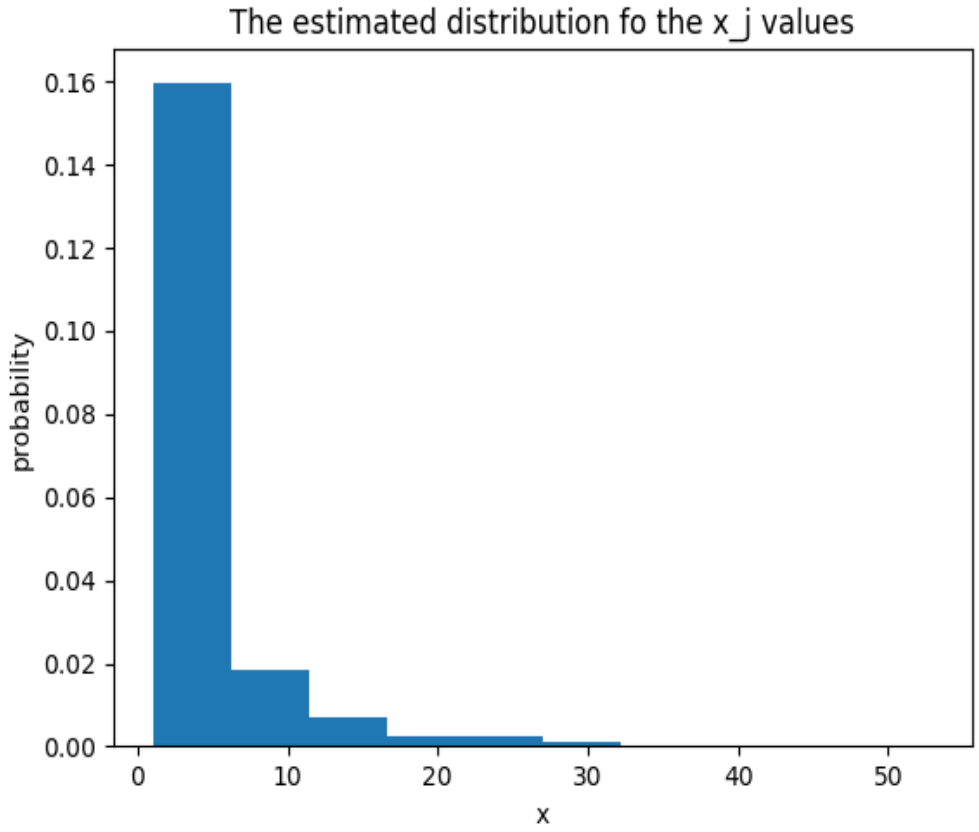

Figure 4.1: *The estimated distribution for the observed x_j values*

As it can be noticed in the Table 4.3 the average number of species observed per sample is 41.91, the median is 27 and the standard deviation is 88.01, also the maximum species observed in a sample was 846 species (the floristic list), the minimum was 2 species and the number of sample occasions was n=91 samples (Table 4.3). Figure 4.2 shows the estimated probability distribution for the observed s_i values during the study.

Average	Median	Standard Deviation	Minimum	Maximum	n
41.91	27	88.01	2	846	91

Table 4.3: *Descriptive measurements for the observed number of species s_i per sample*

https://doi.org/10.26219/heal.aueb.5897

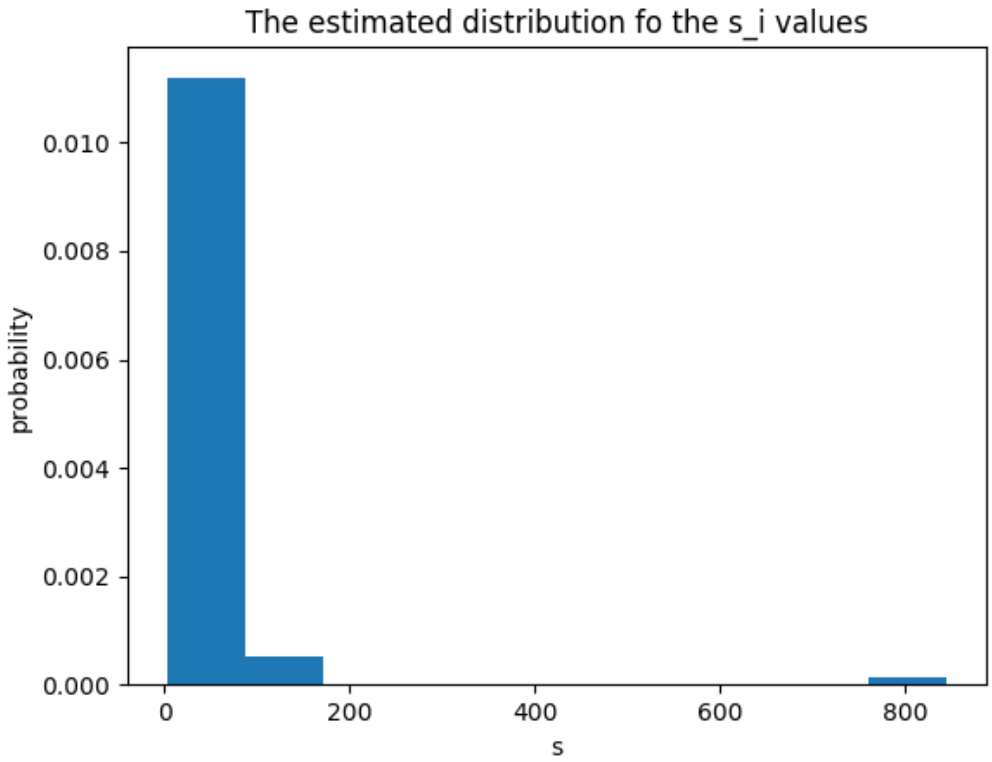

Figure 4.2: *The estimated distribution for the observed* s_i *values*

4.2. Estimation using M_0 model

All the models have distinct relations to each other. The simplest model, M_0, is a special case of all other models. Mathematically, the relations are true because we can assume that some parameters of the more general model are equal to each other to "produce" the simpler model. For example, in Model M_b if all the observing probabilities p_g are assumed to be equal, we have model M_0 (Gary C. White et al. 1982) and the likelihood for model M_0 can be written as:

$$\Pr(f_1, \dots, f_n|F) = \frac{N!}{\prod_i f_i!(N-r)!} \pi_0^{N-r} \prod_{i=1}^{n} \pi_i^{f_i} \quad (5)$$

where $\pi_i = \frac{n!}{(n-i)!i!} p_0^i (1-p_0)^{n-i}$, $i=0,\dots,n$, is the probability that a species is observed i times in n sampling occasions, and p_0 is assumed to be constant between the species and sampling occasions. Thus model M_0 has two parameters to be estimated, N and p_0, however the parameter of interest N can be estimated by profile likelihood and a line search

https://doi.org/10.26219/heal.aueb.5897

The derivative of Eq. 4 with respect to p_0 is given by:

$$\frac{\partial \log[\Pr(f_1, \ldots, f_n|F)]}{\partial p_0} = \frac{\sum_{i=1}^{n} i f_i}{p_0} - \frac{n(N-r) + \sum_{i=1}^{n}(n-i)f_i}{(1-p_0)}$$

and upon setting $\frac{\partial \log[\Pr(f_1, \ldots, f_n|F)]}{\partial p_0}=0$ and solving for p_0 we obtain the maximum likelihood estimator of p_0 given N:

$$\hat{p}_{0,N} = \frac{\sum_{i=1}^{n} i f_i}{n(N-r) + n\sum_{i=1}^{n} f_i}$$

The maximization of the M_0 profile log-likelihood function for N, log $[\Pr(f_1, f_2, \ldots, f_n|\widehat{N}, p_{0,N})]$, with respect to the parameter N, results in the maximum likelihood estimation for N of 979 with standard error of 4.4, a 95% profile likelihood confidence interval of [971.82, 989.44], the value for the maximized logarithmic likelihood is: -1714.4 with AIC and BIC values, 3430 and 3435.6 respectively (Table 4.4).The approximation for the standard error of the parameter is computed as:

$$se(\widehat{N}) \approx \frac{U-L}{4}$$

where, U and L are the upper and lower bounds of the 95% profile likelihood confidence interval, respectively. This approach is used because of the fact that if $\widehat{N}$ followed the normal distribution the 95% confidence interval, based on the normal distribution, for N would be identical to the 95% profile likelihood confidence interval.

Figures 4.3 and 4.4 show the profile log-likelihood function for N and the construction of a 95% profile log-likelihood confidence interval for N from model M_0 respectively. As it can be noticed from the figures model M_0 gives estimation of the total number of different species of plants present in the area very close to the number of the different species observed during the study (r=962), which indicates that the M_0 model lacks some information.

https://doi.org/10.26219/heal.aueb.5897

$\widehat{N}_{ML,M_0}$	Standard error	95% Profile Likelihood C.I.	$log(L_{max})$	AIC	BIC
979.85	4.4	[971.82, 989.44]	-1714.4	3430	3435.6

Table 4.4: *The results derived from the M_0*

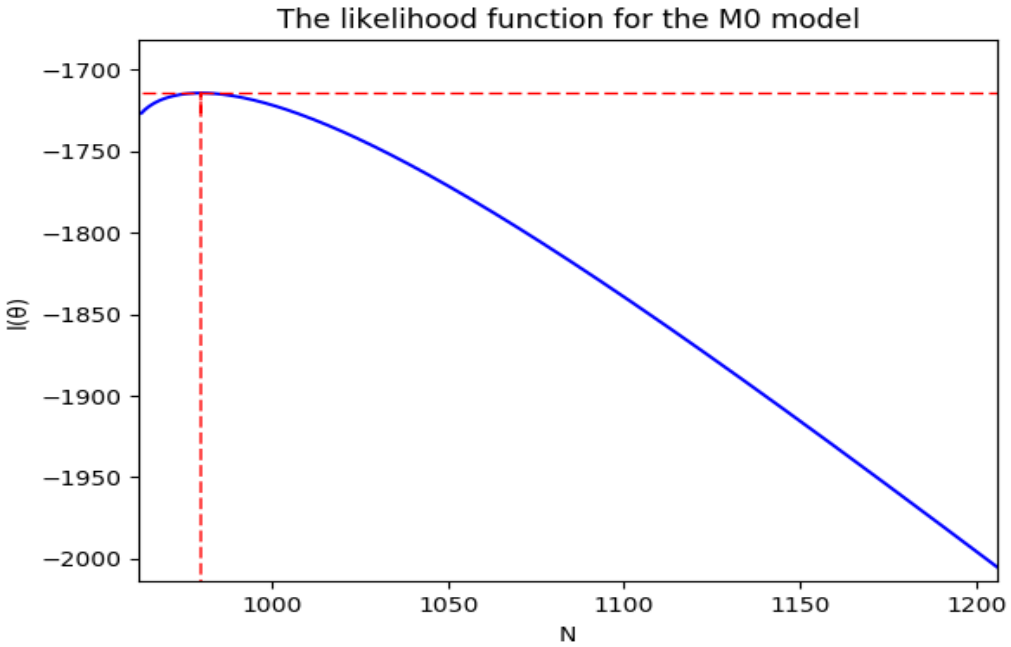

Figure 4.3: *Profile log-likelihood function for N from model M_0*

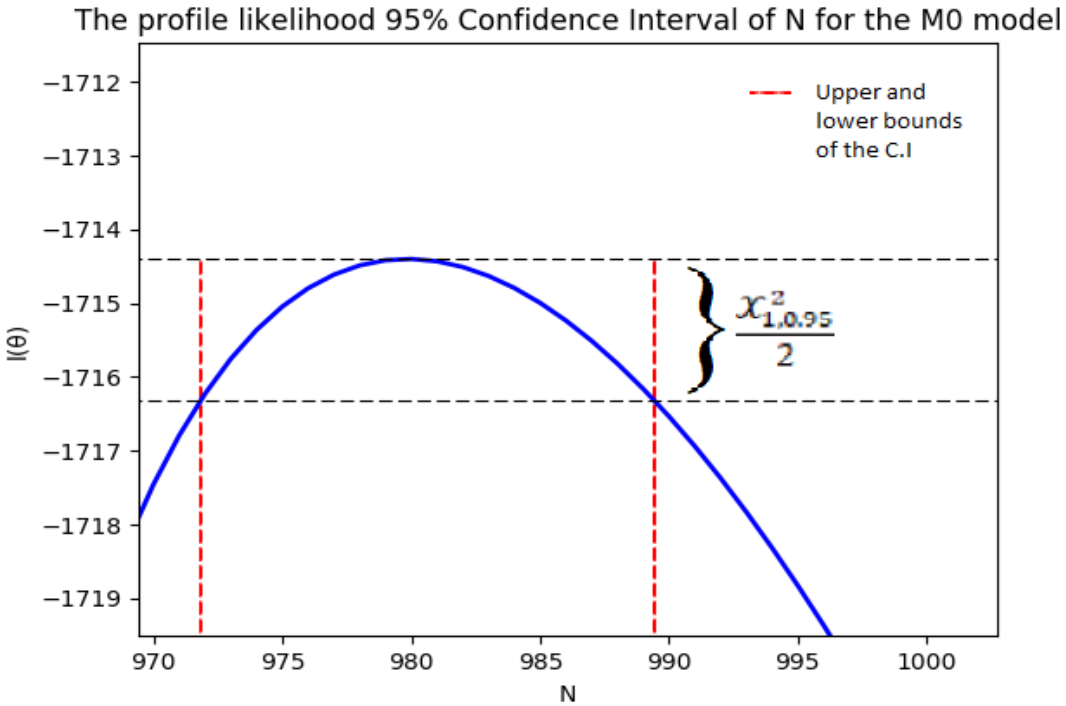

Figure 4.4: *Construction of 95% profile log likelihood confidence interval for N from model M_0*

https://doi.org/10.26219/heal.aueb.5897

4.3. Estimation using $M_{2\text{-}b}$ model

In contrast to model M_0, which assumes a constant observing probability, model $M_{2\text{-}b}$, , assumes that the probability to observe a species varies between two groups of species. Its likelihood can be expressed as:

$$\Pr(f_1, \dots, f_n|F) = \frac{N!}{\prod_i f_i!\,(N-r)!}\pi_0^{N-r}\prod_{i=1}^{n}\pi_i^{f_i}$$

where the probabilities π_i that a species is observed i times in n sampling occasions is now a mixture of two binomial distributions, given by :

$$\pi_i = \sum_{g=1}^{2} w_g \left(\frac{n!}{(n-i)!i!}p_g^i\left(1-p_g\right)^{n-i}\right),\ i=0,..,n.$$

For model $M_{2\text{-}b}$ there are 4 parameters to be estimated: the parameter of interest N which the number of different species present in the area of study, the parameter p_1, which is the probability to observe a species in the first group, and the probability to observe a species i times is given by:

$$\Pr(\mathcal{X}_{observed} = i\,|\,p_1) = \frac{n!}{(n-i)!i!}p_1^i(1-p_1)^{n-i},\ \ i=0,\dots,n,$$

and the parameter p_2 which is the probability to observe a specie in the second group, with probability to observe a species i times given by:

$$\Pr(\mathcal{X}_{observed} = i|\,p_2) = \frac{n!}{(n-i)!i!}p_2^i(1-p_2)^{n-i},\ \ i=0,\dots,n.$$

Also there are weights w_1 and w_2 that should be estimated, which effectively express the proportion of species belonging to the first and second group respectively, subject to $w_1+w_2=1$ therefore the parameters that should be estimated are: N, p_1, p_2 and w_1.

The maximization of the $M_{2\text{-}b}$ logarithmic likelihood with respect to the parameters mentioned above, results in the maximum likelihood estimation for N of 1147.87 with standard error 21.17, a 95% profile likelihood confidence interval of [1108.74, 1193.45], the value for the maximized logarithmic likelihood is: -366 with AIC and BIC values 734.15 and 798.15

https://doi.org/10.26219/heal.aueb.5897

respectively (Table 4.5). Furthermore Table 4.6 shows the maximum likelihood estimations for the parameters p_1, p_2 and w_1 which are found to be 0.14, 0.018 and 0.15 with standard errors 0.003, 0.0006 and 0.01 respectively.

$\widehat{N}_{ML,M_{2-bin}}$	Standard error	95% Profile Likelihood C.I.	$log(L_{max})$	AIC	BIC
1147.87	21.17	[1108.74, 1193.45]	-366	740.15	759.6

Table 4.5: *The results derived from model $M_{2\text{-}b}$*

Parameter	Estimation	Standard Error
p_1	0.14	0.003
p_2	0.018	0.0006
w_1	0.15	0.01

Table 4.6: *Maximum likelihood estimates and estimates of uncertainty for the parameters p_1, p_2 and w*

Figures 4.5 and 4.6 show the profile log-likelihood function for N and the construction of a 95% profile log-likelihood confidence interval for N from model $M_{2\text{-}b}$ respectively. As it can be noticed the standard error for $\widehat{N}$, which describes the uncertainty in the estimate for the number of different species present in the area is larger than that from model M_0.

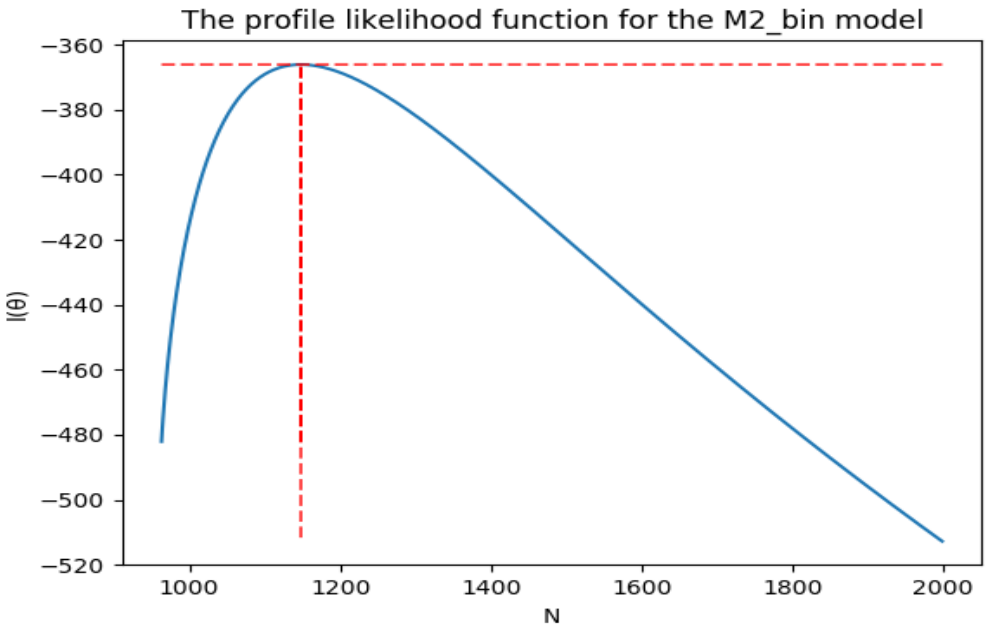

Figure 4.5: *Profile log Likelihood function for N from model $M_{2\text{-}b}$*

https://doi.org/10.26219/heal.aueb.5897

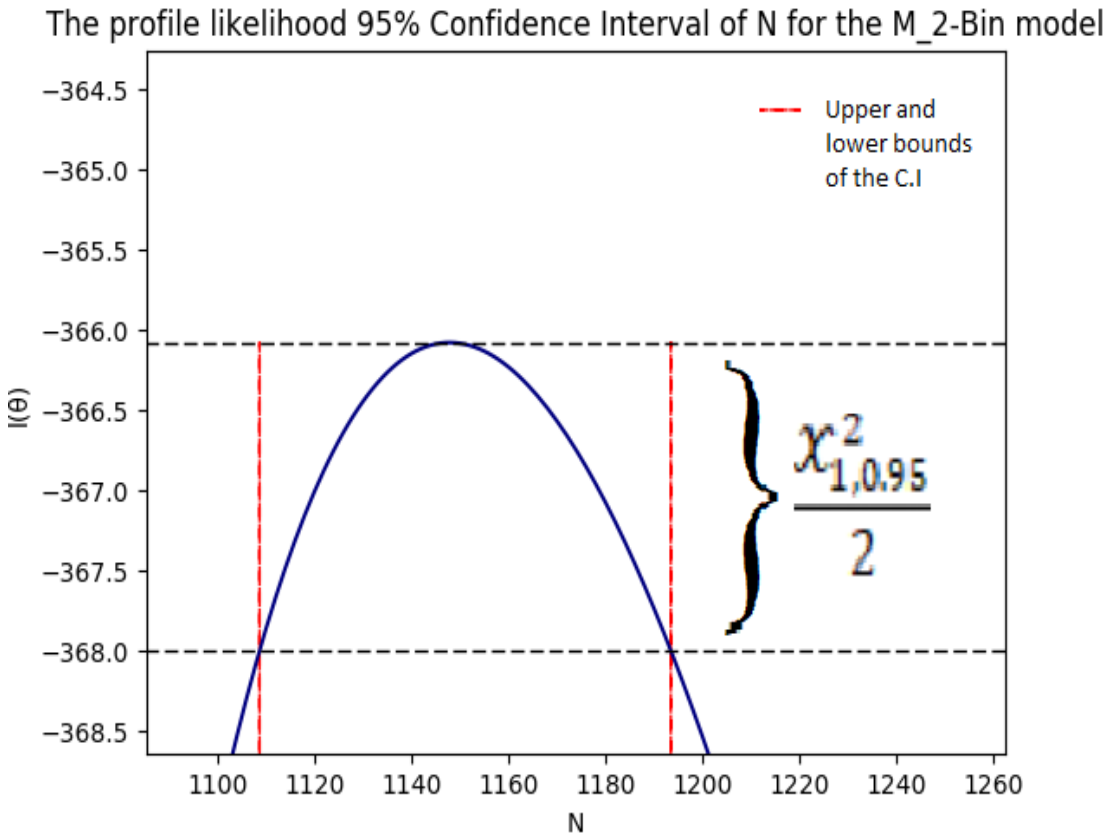

Figure 4.6: *Construction of 95% profile log likelihood confidence interval for N from model $M_{2\text{-}b}$*

4.4. Estimation using $M_{3\text{-}b}$ model

In contrast to model $M_{2\text{-}b}$, model $M_{3\text{-}b}$ assumes that the probability to observe a species varies between three groups of species. Its likelihood can be expressed as:

$$\Pr(f_1, \dots, f_n|F) = \frac{N!}{\prod_i f_i!(N-r)!}\pi_0^{N-r}\prod_{i=1}^{n}\pi_i^{f_i},$$

where the probabilities π_i that a species is observed i times in n sampling occasions is now a mixture of three binomial distributions, given by:

$$\pi_i = \sum_{g=1}^{3} w_g \left(\frac{n!}{(n-i)!i!}p_g^{x_j}(1-p_g)^{n-i}\right), \text{ i=0,…,n.}$$

https://doi.org/10.26219/heal.aueb.5897

For model $M_{3\text{-}b}$ there are 6 parameters to be estimated: the parameter of interest N which the number of different species present in the area of study, the parameter p_1, which is the probability observe a species in the first group and the probability to observe a species i times is given by:

$$\Pr(\mathcal{X}_{observed} = i | p_1) = \frac{n!}{(n-i)!i!} p_1^i (1-p_1)^{n-i}, \quad i=0,\ldots,n,$$

the parameter p_2 which is the probability to observe a specie in the second group, with probability to observe a species i times given by:

$$\Pr(\mathcal{X}_{observed} = i | p_2) = \frac{n!}{(n-i)!i!} p_2^i (1-p_2)^{n-i}, \quad i=0,\ldots,n,$$

and the parameter p_3 which is the probability to observe a specie in the third group, with probability to observe a species i times given by:

$$\Pr(\mathcal{X}_{observed} = i | p_3) = \frac{n!}{(n-i)!i!} p_3^i (1-p_3)^{n-i}, \quad i=0,\ldots,n.$$

Also there are weights w_1, w_2 and w_3 that should be estimated, which effectively express the proportion of species belonging to the first, second and third group respectively, subject to $w_1+w_2+w_3=1$ therefore the parameters that should be estimated are: N, p_1, p_2, p_3, w_1 and w_2.

The maximization of the $M_{3\text{-}b}$ logarithmic likelihood with respect to the parameter N, results in the maximum likelihood estimation for N of 1282 with standard error of 36.4, a 95% profile likelihood confidence interval of [1216.2, 1361.7], the value for the maximized logarithmic likelihood is: -144.2 with AIC and BIC values 300.45 and 354.45 respectively (Table 4.7). Furthermore Table 4.8 shows the maximum likelihood estimations for the parameters p_1, p_2, p_3, w_1 and w_2 which are found to be 0.27, 0.087, 0.013, 0.027 and 0.16 with standard errors 0.008, 0.0027, 0.0006, 0.0047 and 0.011 respectively.

https://doi.org/10.26219/heal.aueb.5897

$\widehat{N}_{ML,M_{3-bin}}$	Standard error	95% Profile Likelihood C.I.	$log(L_{max})$	AIC	BIC
1282	36.4	[1216.2, 1361.7]	-144.2	300.45	329.6

Table 4.7: *The results derived from model $M_{3\text{-}b}$*

Parameter	Estimation	Standard Error
p_1	0.27	0.008
p_2	0.087	0.0027
p_3	0.013	0.0006
w_1	0.027	0.0047
w_2	0.16	0.011

Table 4.8: *Maximum likelihood estimates and estimates of uncertainty for the parameters p_1, p_2, p_3, w_1 and w_2*

Figures 4.7 and 4.8 show the profile log-likelihood function for *N* and the construction of a 95% profile log-likelihood confidence interval for *N* from model $M_{3\text{-}b}$ respectively.

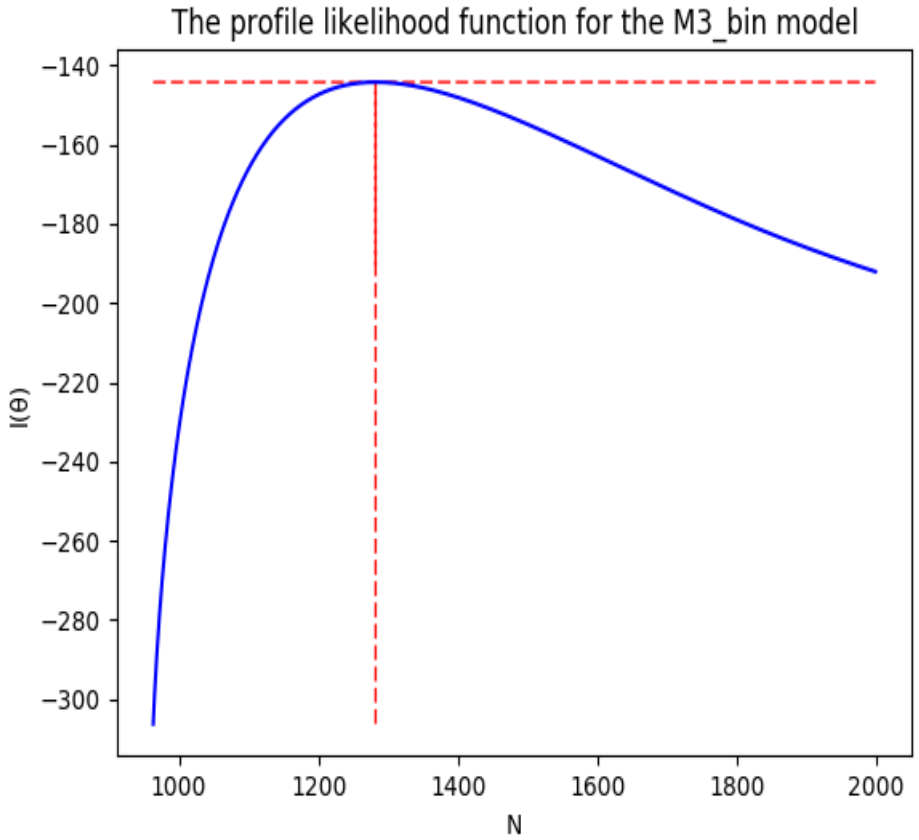

Figure 4.7: *Profile log-likelihood function for N from model $M_{3\text{-}b}$*

https://doi.org/10.26219/heal.aueb.5897

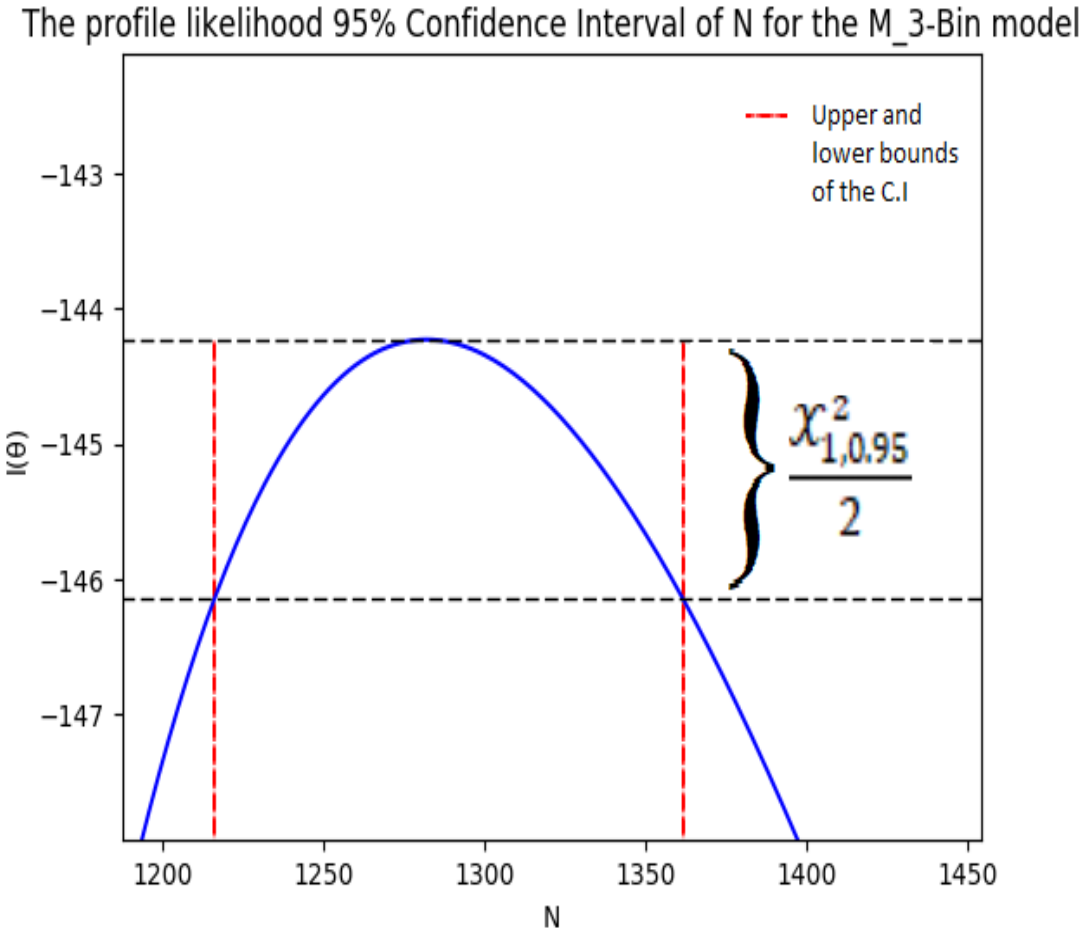

Figure 4.8: *Construction of 95% profile log likelihood confidence interval for N from model* $M_{3\text{-}b}$

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 5.

BOOTSTRAP IMPLIMENTATION

5.1 The Bootstrap algorithm

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of re-sampling methods.

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution function of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of re samples from the observed dataset (and of equal size to the observed dataset).

It may also be used for constructing hypothesis tests. It is often used as an alternative to statistical inference based on the assumption of a parametric model when that assumption is in doubt, or where parametric inference is impossible or requires complicated formulas for the calculation of standard errors.

For capture-recapture the observe data $\mathcal{X} = [x_1, x_2, \dots, x_r, \underbrace{0, \dots, 0}_{\widehat{N}_{Ml} - r}]$ which is the empirical frequencies of the observed data. Where, $x_j = \sum_{i=1}^{i=n} I_{j,i}$, r is the number of different species observed during the study and $\widehat{N}_{Ml}$ is the estimation of the number of different species present in the area according to

https://doi.org/10.26219/heal.aueb.5897

a model. Having the above data, the algorithm for the bootstrap technique has the beneath steps:

- Generate the vector $\mathcal{X}_{bootstrap}$ by sampling with replacement from the vector $\mathcal{X} = [x_1, x_2, \dots, x_r, \underbrace{0, \dots, 0}_{\widehat{N}_{Ml}-r}]$, where $\widehat{N}_{Ml}$ is the maximum likelihood estimation of the number of different species present in the area according to a model.

- Get the $\mathcal{X}^{+}_{bootstrap}$ by keeping only the values which are greater than zero from the $\mathcal{X}_{bootstrap}$ sample.

- Create the $f_{bootstrap}$ vector as:

$$f_{bootstrap} = [f_{1,bootstrap}, \dots, f_{i,bootstrap}, \dots, f_{n,bootstrap}]$$

where, $f_{i,bootstrap} = \sum_{j=1}^{r_{bootstrap}} I(x^{+}_{j,bootstrap} = i)$, $i=1, 2, \dots, n$
and $r_{bootstrap}$ is the length of the vector $\mathcal{X}^{+}_{bootstrap}$

- Estimate the number of species $\widehat{N}_{Ml,boostrap}$ from frequency vector $f_{bootstrap}$, assuming a model

- Doing the above procedure for *R* replications a good estimation for the probability distribution of the estimation of the total number of different species present in the area, *N*, can be obtained. Having this estimation of the probability distribution of the different species present in the area the standard error and the percentile *(1-α)%* confidence can be obtained for the estimation of *N*.

https://doi.org/10.26219/heal.aueb.5897

5.2. Bootstrap for model M_0

Figure 5.1 shows the estimated probability distribution of $\hat{N}$ from model M_0, using the bootstrap algorithm with R=10000 replications. Table 4.1 provides the standard error, the average value and the 95% percentile confidence interval for the parameter estimation of N based on the bootstrap replications for the model M_0.

As it can be seen from Table 5.1 the bootstrap procedure results in an, average value of 980.14, standard error of 5.54 and a 95% percentile confidence interval of [969.23, 991.17].

Average	Standard error	95% Percentile Bootstrap C.I.
980.14	5.54	[969.23, 991.17]

Table 5.1: *Summary results from the implementation of the bootstrap procedure for the M_0 model*

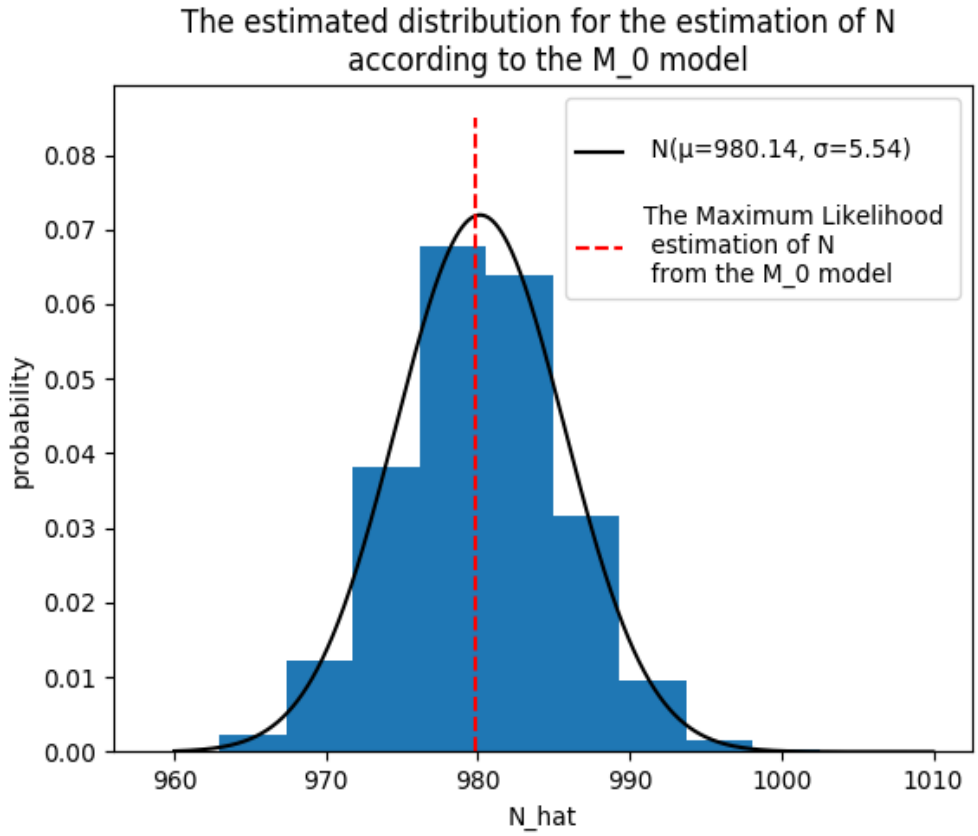

Figure 5.1: *Histogram of the $\hat{N}_{ML,M_0}$ from 10000 bootstrap replications together with a normal distribution based on the mean and standard deviation of the bootstrap estimates. The value $\hat{N}_{ML,M_0}$=979.14 is also shown*

https://doi.org/10.26219/heal.aueb.5897

5.3. Bootstrap for model $M_{2\text{-}b}$

Figure 5.2 shows the estimated probability distribution of $\hat{N}$ from model $M_{2\text{-}b}$, using the bootstrap algorithm with R=1000 replications. Table 4.2 provides the standard error, the average value and the 95% percentile confidence interval for the parameter estimation of N based on the bootstrap replications for the model $M_{2\text{-}b}$.

As it can be seen from Table 5.2 the bootstrap procedure results in an, average value of 1153.11, standard error of 33 and a 95% percentile confidence interval of [1093.91, 1221.7].

Average	Standard error	95% Percentile Bootstrap C.I.
1153.11	33	[1093.91, 1221.7]

Table 5.2: *Summary results from the implementation of the bootstrap procedure for the $M_{2\text{-}b}$ model*

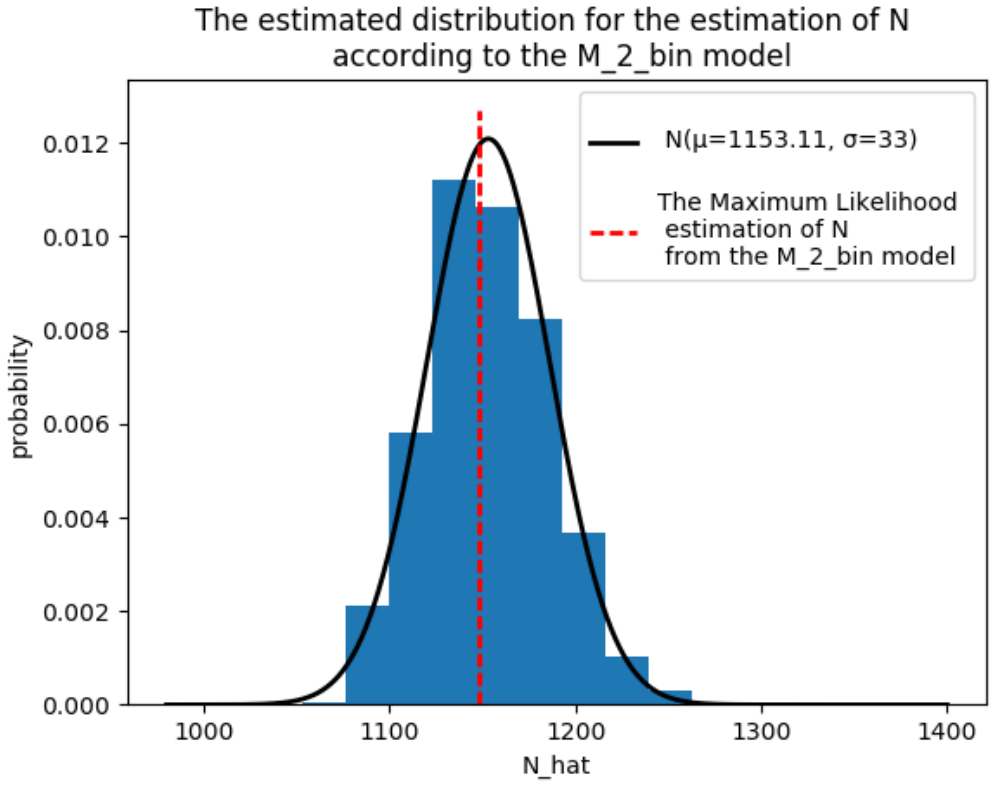

Figure 5.2: *Histogram of the $\hat{N}_{ML,M_{2-b}}$ from 1000 bootstrap replications together with a normal distribution based on the mean and standard deviation of the bootstrap estimates. The value $\hat{N}_{ML,M_{2-b}}$=1147.87 is also shown*

https://doi.org/10.26219/heal.aueb.5897

5.4. Bootstrap for model $M_{3\text{-}b}$

Figure 5.3 shows the estimated probability distribution of $\hat{N}$ from model $M_{3\text{-}b}$, using the bootstrap algorithm with R=1000 replications. Table 4.3 provides the standard error, the average value and the 95% percentile confidence interval for the parameter estimation of N based on the bootstrap replications for the model $M_{3\text{-}b}$.

As it can be seen from Table 5.3 the bootstrap procedure results in an, average value of 1291.51, standard error of 51.22 and a 95% percentile confidence interval of [1201.83, 1399].

Average	Standard error	95% Percentile Bootstrap C.I.
1291.51	51.22	[1201.83, 1399]

Table 5.3: *Summary results from the implementation of the bootstrap procedure for the $M_{3\text{-}b}$ model*

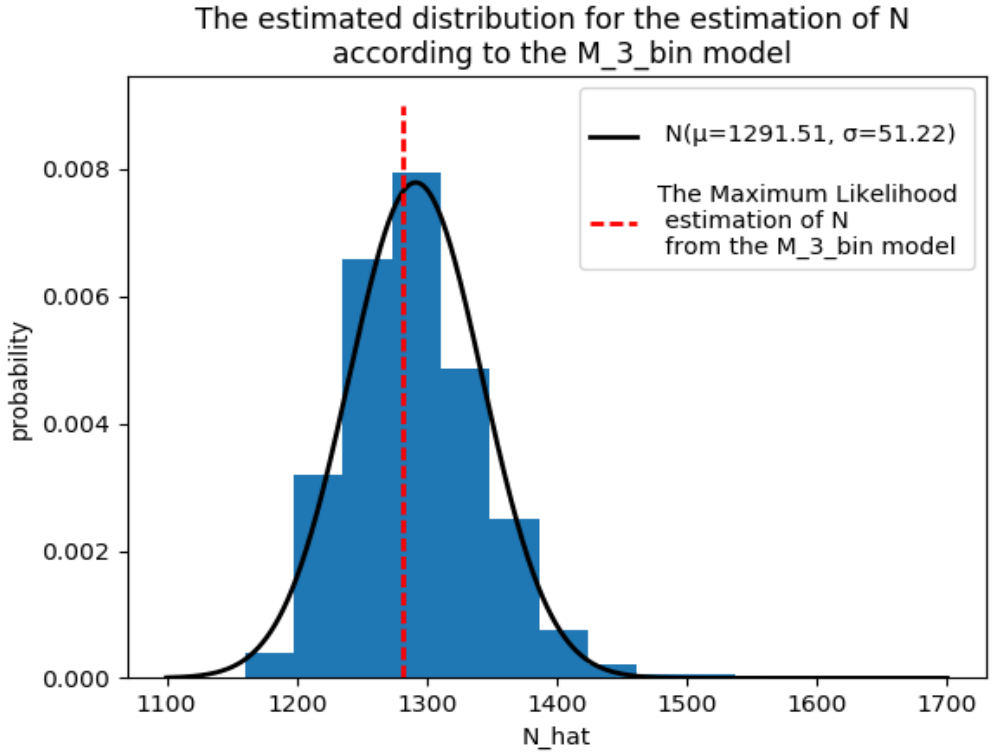

Figure 5.3: *Histogram of the $\hat{N}_{ML,M_{3-b}}$ from 1000 bootstrap replications together with a normal distribution based on the mean and standard deviation of the bootstrap estimates. The value $\hat{N}_{ML,M_{3-b}}$=1282 is also shown*

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 6.

MONTE-CARLO IMPLIMENTATION

6.1. The Monte-Carlo algorithm

The Monte-Carlo technique is similar to the bootstrap technique, as it was described in the section, with the only difference is that now instead of sampling with replacement directly from the vector $\mathcal{X}$, samples will be generated from the parametric distributions *f(p)* assumed according to each model. For each sample, respectively to the bootstrap, $f_{bootstrap}$ frequency vectors, $f_{Monte\text{-}Carlo}$ frequency vectors will be created based on the generated from the parametric distribution fp, vector $\mathcal{X}^{+}_{Monte-Carlo}$. Finally using the $f_{Monte\text{-}Carlo}$ frequency vector, an estimation for *N* will be computed, assuming one of the models. This procedure will be performed for *R* replications from which a distribution of the estimation of *N*, according to each of the models, will be obtained.

6.2. Monte-Carlo for model M_0

Figure 6.1 shows the estimated probability distribution of $\widehat{N}$ from model M_0, using the Monte-Carlo algorithm with R=10000 replications. Table 5.1 provides the standard error, the average value and the 95% percentile confidence interval for the parameter estimation of *N* based on the Monte-Carlo replications for the model M_0.

As it can be seen from Table 6.1 the Monte-Carlo procedure results in an, average value of 979.8, standard error of 1.2 and a 95% percentile confidence interval of [977.5, 982.2].

Average	Standard error	95% Percentile Bootstrap C.I.
979.8	1.2	[977.5, 982.2]

Table 6.1: *Summary results from the implementation of the Monte-Carlo procedure for the M_0 model*

https://doi.org/10.26219/heal.aueb.5897

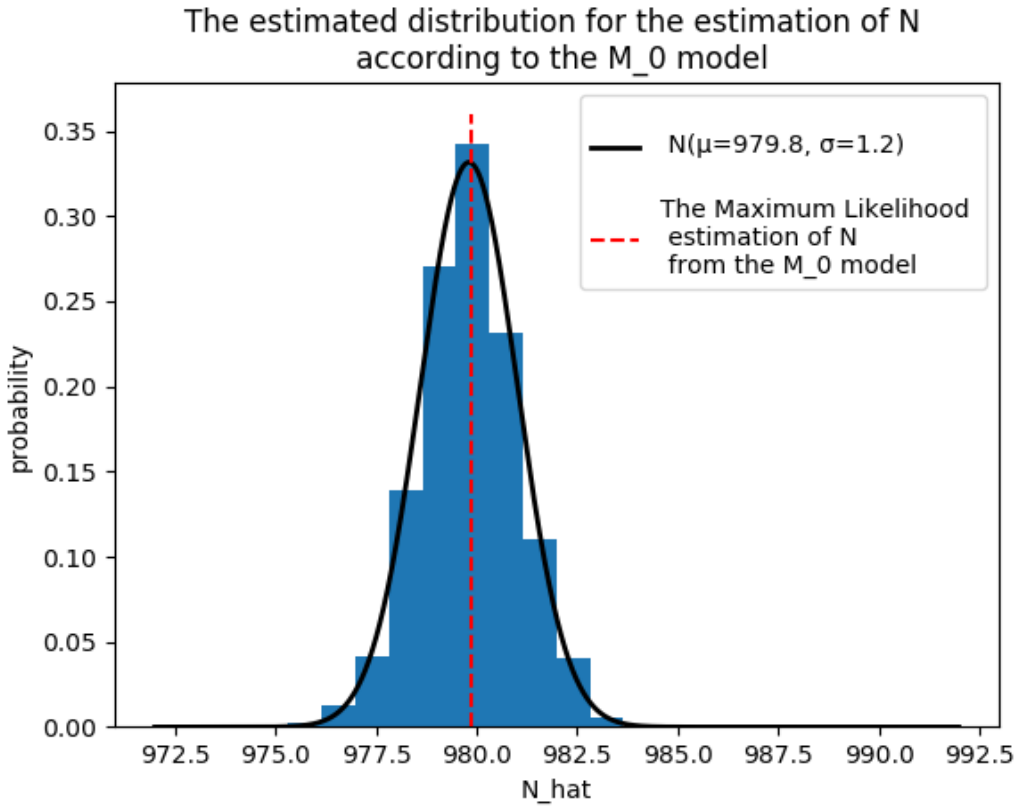

Figure 6.1: *Histogram of the $\widehat{N}_{ML,M_0}$ from 10000 Monte-Carlo replications together with a normal distribution based on the mean and standard deviation of the Monte-Carlo estimates. The value $\widehat{N}_{ML,M_0}$=979.14 is also shown*

6.3. Monte-Carlo for model $M_{2\text{-}b}$

Figure 6.2 shows the estimated probability distribution of $\widehat{N}$ from model $M_{2\text{-}b}$, using the Monte-Carlo algorithm with R=1000 replications. Table 5.2 provides the standard error, the average value and the 95% percentile confidence interval for the parameter estimation of N based on the Monte-Carlo replications for the model $M_{2\text{-}b}$.

As it can be seen from Table 6.2 the Monte-Carlo procedure results in an, average value of 1148, standard error of 20.38 and a 95% percentile confidence interval of [1108, 1188.7].

https://doi.org/10.26219/heal.aueb.5897

Average	Standard error	95% Percentile Bootstrap C.I.
1148	20.38	[1108, 1188.7]

Table 6.2: *Summary results from the implementation of the Monte-Carlo procedure for the $M_{2\text{-}b}$ model*

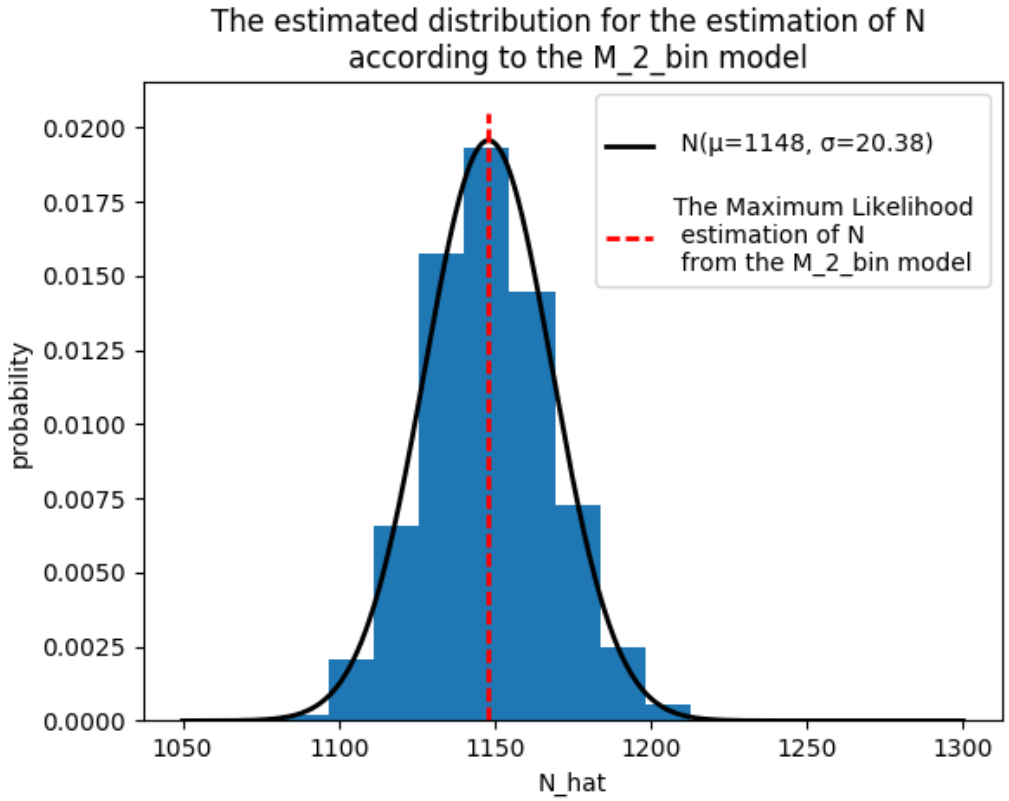

Figure 6.2: *Histogram of the $\widehat{N}_{ML,M_{2-b}}$ from 1000 Monte-Carlo replications together with a normal distribution based on the mean and standard deviation of the Monte-Carlo estimates. The value $\widehat{N}_{ML,M_{2-b}}$=1147.87 is also shown*

6.4. Monte-Carlo for model $M_{3\text{-}b}$

Figure 6.3 shows the estimated probability distribution of $\widehat{N}$ from model $M_{3\text{-}b}$, using the Monte-Carlo algorithm with R=1000 replications. Table 5.3 provides the standard error, the average value and the 95% percentile confidence interval for the parameter estimation of N based on the Monte-Carlo replications for the model $M_{3\text{-}b}$.

As it can be seen from Table 6.3 the Monte-Carlo procedure results in an, average value of 1283, standard error of 34 and a 95% percentile confidence interval of [1219.75, 1350.32].

https://doi.org/10.26219/heal.aueb.5897

Average	Standard error	95% Percentile Bootstrap C.I.
1283	34	[1219.75, 1350.32]

Table 6.3: *Summary results from the implementation of the Monte-Carlo procedure for the $M_{3\text{-}b}$ model*

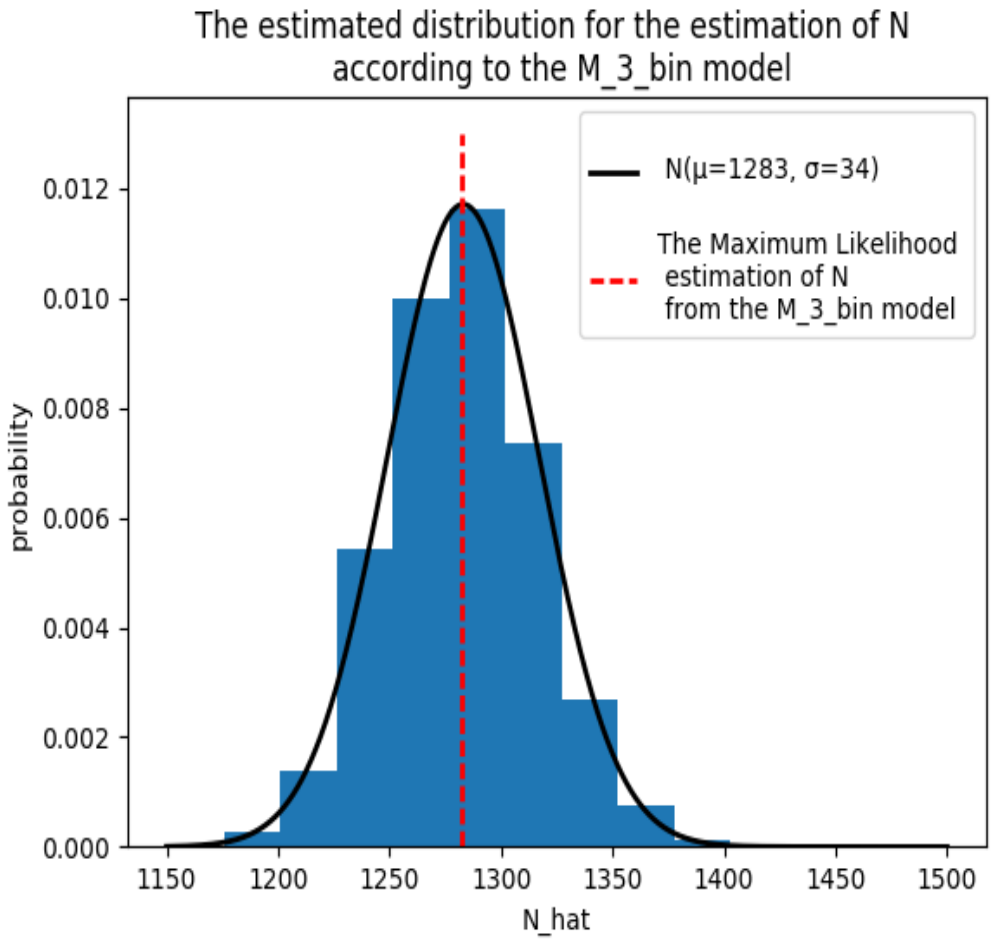

Figure 6.3: *Histogram of the $\widehat{N}_{ML,M_{3-b}}$ from 1000 Monte-Carlo replications together with a normal distribution based on the mean and standard deviation of the Monte-Carlo estimates. The value $\widehat{N}_{ML,M_{3-b}}$=1282 is also shown*

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 7.

MODELS WITH *G*>3

As it can be noticed from Tables 4.5 and 4.7 the AIC and BIC values decrease significantly from model $M_{2\text{-}b}$ (AIC=740.15, BIC=759.6) to model $M_{3\text{-}b}$ model (AIC=300.45, BIC=329.6). In this section additional components are added to the mixture of distributions:

$$\pi_i = \sum_{g=1}^{G} w_g \left(\frac{n!}{(n-i)!i!} p_g^i (1-p_g)^{n-i} \right),\ i=0,\dots,n.$$

Table 7.1 gives the estimations of N, the standard errors of the estimations as well as the AIC and BIC values for G=4, 5 and 6. Also Table 7.2 provides the estimations and the standard errors for the parameters p_g and w_g, for g=4, 5 and 6. Since the computational complexity increases and the derivation of the profile likelihood for each of the models becomes a really difficult task, the standard errors and the confidence intervals for the models with G=5 and G=6 are computed via bootstrap simulations as described in the previous chapter, while for the model with G=4, Monte-Carlo simulations were used.

Model	$\widehat{N}_{ML}$	Standard error	95% Confidence Interval	$log(L_{max})_n$	AIC	BIC
$M_{4\text{-}b}$	1349.4	53	[1332, 1540]	-106.7	229.43	268.3
$M_{5\text{-}b}$	1634.3	170.4	[1372.5, 2025]	-87.3	194.6	243.3
$M_{6\text{-}b}$	1723.4	210.8	[1396, 2176]	-82.1	188.2	246.6

Table 7.1: *The estimation of N for each of the models*

https://doi.org/10.26219/heal.aueb.5897

Model								
$M_{4\text{-}b}$			$M_{5\text{-}b}$			$M_{6\text{-}b}$		
Parameter	Estimation	S.E.	Parameter	Estimation	S.E.	Parameter	Estimation	S.E.
p_1	0.18	0.007	p_1	0.25	0.02	p_1	0.34	0.03
p_2	0.4	0.02	p_2	0.11	0.012	p_2	0.1	0.009
p_3	0.07	0.003	p_3	0.04	0.007	p_3	0.035	0.006
p_4	0.01	0.0008	p_4	0.48	0.08	p_4	0.54	0.06
w_1	0.043	0.005	p_5	0.006	0.001	p_5	0.2	0.02
w_2	0.007	0.002	w_1	0.018	0.003	p_6	0.005	0.001
w_3	0.16	0.01	w_2	0.07	0.01	w_1	0.004	0.002
			w_3	0.17	0.015	w_2	0.076	0.009
			w_4	0.002	0.002	w_3	0.18	0.02
						w_4	0.001	0.002
						w_5	0.02	0.004

Table 7.2: *The estimation of the parameters p_g and w_g and their standard errors for each model*

We can see in table 6.1 the AIC value keeps decreasing, therefore one more binomial is added to the π_i making the number of binomials G=7 and π_i is expressed as:

$$\pi_i = \sum_{g=1}^{7} w_g \left(\frac{n!}{(n-i)!i!} p_g^i (1-p_g)^{n-i} \right),\ i=0,\ldots,n.$$

The model with a mixture of 7 binomials gives maximum likelihood estimations for N of $\widehat{N}_{ML}$= 1990.2 the maximized logarithmic likelihood is: -1429.15 with the values of AIC and BIC of 2886.3 and 2954.5 respectively (Table 7.3), the likelihood value is lower than the likelihood values from models with G=5 and 6, therefore the results for the model with G=7 are not reliable. As it can be noticed the AIC value decrease is very low from G=5 to G=6, while the BIC values start increasing for G>5, which indicates that the models with 5 and 6 binomial distributions are the best fitting models according to the BIC and AIC respectively. Figure 7.1 plots the AIC and BIC curves for each of the models.

Model	$\widehat{N}_{ML}$	$log(L_{max})_{max}$	AIC	BIC
$M_{7\text{-bin}}$	1990.2	-1429.15	2886.3	2954.5

Table 7.3: *Estimation of N from model $M_{7\text{-}b}$*

https://doi.org/10.26219/heal.aueb.5897

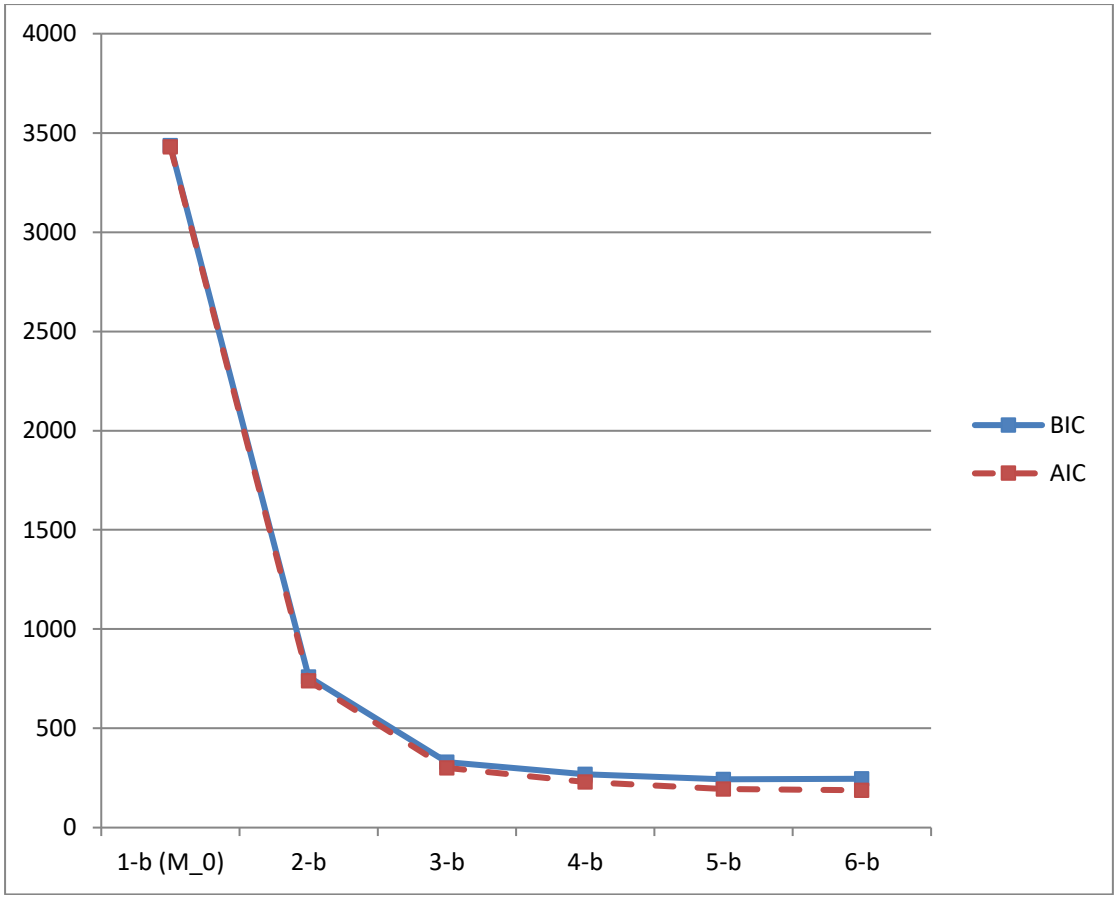

Figure 7.1: *BIC and the AIC values for models ($M_{1\text{-}b}$=M_0)*

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 8.

MODEL AVERAGING

In order to obtain information from all the models that have been used to estimate the total number of different species present in the area of study, model averaging will be used; for this purpose weights w_m will be computed based on AIC values of the m^{th} model as:

$$w_m = \frac{e^{-0.5\Delta_m}}{\sum_{m=1}^{M} e^{-0.5\Delta_m}} \ ,$$

where:

- M is the number of different models that were used
- $\Delta_m = AIC_m - AIC_{min}$

Table 8.1 shows the AIC weights of each model based on their AIC values where it can be noticed that the $M_{6\text{-}b}$ has the largest weight (w =0.96) and the weights for all the other models are significantly lower which means that effectively the only model which can be included in the estimation of N, according to the AIC weights, is the $M_{6\text{-}b}$ model. Table 8.2 shows the estimation of N from the best fitting model based on AIC value as well the estimation of N from the best fitting model based on BIC value.

Model	AIC weight
M_0	<0.0001
$M_{2\text{-}b}$	<0.0001
$M_{3\text{-}b}$	<0.0001
$M_{4\text{-}b}$	<0.0001
$M_{5\text{-}b}$	0.04
$M_{6\text{-}b}$	0.96
$M_{7\text{-}b}$	<0.0001

Table 8.1: *The AIC weights for each model*

https://doi.org/10.26219/heal.aueb.5897

Criteria	Estimation	Standard Error	95% Confidence Interval
AIC	1723.4	210.8	[1396.9, 2176]
BIC	1634.3	170.4	[1372.5, 2025]

Table 8.2: *The estimation of N as well as the standard error and a 95% c.i. based on the lowest AIC and BIC values*

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 9.

THE M_t MODEL

Model M_t assumes that the probability to observe a species is constant between the species but varies from sample to sample. The likelihood can be expressed as (see Eq. 2):

$$\Pr(\{\alpha_\omega\}) = \frac{N!}{\prod_\omega \alpha_\omega!(N-r)!} \prod_{i=1}^{n} p_i^{s_i} q_i^{N-s_i} \text{ , (3)}$$

where p_i is the probability to observe a species in sample i, $i=1,..,n$. There are thus n+1 parameters in the model (N, $p_1,\ldots,p_n$), and estimation can be difficult when n is large as in Maremma Regional Park study, where n=91. However we can reduce the dimension of the numerical search to a scalar search for N as follows. Setting the derivative of the log-likelihood with respect to p_i equal to zero:

$$\frac{\partial \log[\Pr(\{\alpha_\omega\})]}{\partial p_i} = \frac{s_i}{p_i} - \frac{N-s_i}{1-p_i} = 0,\ i=1,..,n$$

and solving with respect to p_i $\frac{\partial \log[\Pr(\{\alpha_\omega\})]}{\partial p_i}$ we obtain the maximum likelihood estimator of p given N:

$$\hat{p}_i = \frac{s_i}{N}.$$

Substituting $\hat{p}_i$ for p_i in Eq. 3 results in the profile likelihood for N given by:

$$\Pr(\{\alpha_\omega\}) = \frac{N!}{\prod_\omega \alpha_\omega!\,(N-r)!} \prod_{i=1}^{n} \left(\frac{s_i}{N}\right)^{s_i} \left(1-\frac{s_i}{N}\right)^{N-s_i}$$

and the only parameter that remains to be estimated is N. Table 9.1 shows that the maximum likelihood estimation of N according to the M_t model is 968 with a standard error of 2.25 and a 95% confidence interval of [963, 972]. As it can be noticed the estimation of N is very close to the number of observed

https://doi.org/10.26219/heal.aueb.5897

species (r=962) which means that model M_t does not offer much of information about the number of different species present in the area of study. An explanation to the failure of the model M_t can be the fact that this model assumes different observing/capture probability per sample, and this might be not the most sensible assumption for the given study, since whether a plant of a species j will be observed depends only by the researchers who are investigating a given sample/plot i; additionally the fact that the samples/plots were created during the same period of year, leads to consider the model M_t not appropriate.

$\widehat{N}_{ML}$	Standard error	95% Confidence Interval
968	2.25	[963, 972]

Table 9.1: *The estimation of N, the standard error and a 95% c.i. from the M_t model*

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 10.

THE BETA-BINOMIAL MODEL

The beta-binomial model is a heterogeneity model like the models in sections 3 and 6, but the model assumes that the number of binomial distributions in the π_i mixture is infinite since $0<p<1$, a natural distribution to describe the heterogeneity in p would be the $f(p)=Beta(\alpha,\beta)$, and the beta-binomial model can now be expressed as:

$$\Pr(f_1, \dots, f_n|F) = \frac{N!}{\prod_i f_i!\,(N-r)\,!}\pi_0^{N-r}\prod_{i=1}^{n}\pi_i^{f_i}$$

where π_i are the probabilities which are defined as:

$$\pi_i = \int_0^1 \frac{n!}{(n-i)!\,i!}p^i(1-p)^{n-i}dF(p)$$

$$= \int_0^1 \frac{n!}{Be(a,\beta)(n-i)!i!}p^i(1-p)^{n-i}p^{a-1}(1-p)^{\beta-1}dp$$

$$= \frac{n!}{Be(a,\beta)(n-i)!\,i!}\int_0^1 p^i(1-p)^{n-i}p^{a-1}(1-p)^{\beta-1}\,dp$$

$$= \frac{n!}{Be(a,\beta)(n-i)!\,i!}\int_0^1 p^{i+a-1}(1-p)^{n-i+\beta-1}\,dp$$

$$= \frac{n!\,Be(i+a,n-i+\beta)}{Be(a,\beta)(n-i)!\,i!}$$

The maximization of the beta-binomial likelihood with respect to the parameters N, α and β is problematic for the Maremma Park study, resulting in $\widehat{N} \to \infty$ An explanation to this might be the fact that the estimation of N is very large according to the beta-binomial model might be the fact that the fitted beta distribution has a mode at zero. This is simply because the fitted

https://doi.org/10.26219/heal.aueb.5897

model then implies high uncertainty regarding knowledge of *N*, as a consequence of low capture probabilities (McCrea and Morgan, 2015, p45) also large estimates of *N* can arise when an appreciable amount of capture probability is located close to the origin (McCrea and Morgan, 2015, p39). For this reason the beta-binomial was not investigated further, or included in the Model-averaging in Chapter 8.

https://doi.org/10.26219/heal.aueb.5897

CHAPTER 11.

CONCLUSION/FURTHER RESEARCH

The main conclusion that can be made is that, for the given data, the most appropriate approach in order to estimate the number of different species in the whole area, which leads to sensible or/and informative results, is the usage of the finite mixture of distributions. The best models according to the finite mixture of binomial distributions are the one with a mixture of five binomials (according to the BIC) and the one with a mixture of six binomials (according to the AIC), with the estimations of the total number of different species derived from the two models to be at 1634 and 1723 respectively.

The further research that can be performed on the available data is based on the fact that a fairly good indication about what is going on under the ground, can be what is going on over the ground; by transposing the table of the data (that is placing the plots on the rows and the species on the columns), clustering or unsupervised learning can be performed, which will split the plots to p classes and based on this classes each new plot, with a given sequence of species, could be classified on each of the classes. Additionally, if there was information of interest about each plot's underground, supervised learning could be performed which could classify each new plot with a given sequence of species on its surface as regard the information of interest. Another direction for further research based on the stochastic modeling in order to estimate populations, could be the outer space, for example if a new solar system is discovered and we want to estimate the number of different planets present in this newly discovered solar system, we implement the stochastic modeling; in this case the "plots" are the different parts of the solar system and the "species" are the different planets observed on these parts of the solar system.

https://doi.org/10.26219/heal.aueb.5897

REFERENCES

Bailey (1969)

Beukema and de Vos (1974)

Burnham and Overton, 1969

Burnham and Overton, 1978

Burnham(1972)

Carothers (1973a)

Carothers, 1973b

Chao, 1987

Chiarucci et al (2018)

Crowcroft and Jeffers (1961)

Darroch 1958

Edwards and Eberhardt(1967)

Gary C. White et al. 1982

Gilbert, 1973

Gliwicz (1970)

H. Stryhn, J. Christensen

Huber (1962)

Jensen (1975)

Manly, 1970

McCrea and Morgan, 2015, p39

McCrea and Morgan, 2015, p45

Montgomery (1979)

Otis et al., 1978

Otis et al., 1978:33-37

Seber, 1973, Chap. 3

Seber, 1980

Seber, 1982

Skalski and Robson, 1992

Tanaka (1956)

Young et al. (1952)

Καρανδεινός, Μ.Γ. (2007). Ποσοτικές Οικολογικές Μέθοδοι. Παν. Εκδόσεις Κρήτης

https://doi.org/10.26219/heal.aueb.5897