UNIVERSITY OF BERGEN
Faculty of Mathematics and Natural Sciences

Close-Kin Mark-Recapture Models

Brage Førland

Master's thesis in statistics

Department of Mathematics

University of Bergen

June 3, 2019

Summary

Close-Kin Mark-Recapture (CKMR) is a recent extension of the ordinary mark–recapture methods used to estimate animal abundance and other population parameters. Where ordinary mark–recapture only consider the subsequent identification of the same animal a recapture, CKMR expands this by also viewing the genetic identification of a relatives as a recapture. One of the challenges of CKMR models compared to ordinary mark–recapture is that the recapture probabilities are tightly coupled to the life histories of the animals in questions.

This thesis contains three different contributions to the CKMR literature. Firstly I develop a CKMR estimator for age structured populations, presented in Ruzzante (2019)[1]. Secondly, I develop theoretical background for half sibling CKMR analysis, and apply kin analysis to data from the River Etne. Thirdly, it expands on the results from Skaug (2017)[2] and derives several new results for the case where age of both parent and offspring is unknown.

The first part contains the method development of a parent–offspring CKMR model for brook trout populations, electrofished yearly in the period 2013-2018. I here develop a moment estimator for population size for an age structured model, related to the Lincoln–Petersen estimator. The estimator is applied under two different population assumptions, stable age structure, and variable recruitment and representative sampling. Special focus is on the small population situation, where large sample approximations used in previous CKMR studies cannot be assumed. A small sample bias correction for the estimator is developed and validated using parametric bootstrap simulations. Using the perspective that the parent marks the offspring instead of the commonly used offspring marks juvenile, a simple and general form of the estimator is derived. Viewing offspring as the marked part of the population also leads to an expression for the variance of the expected number of parent–offspring pairs in a sample, which is found to be less than the Poisson variance unless fecundity is very overdispersed.

The second part contains theoretical background and model development for half sibling CKMR analysis, to examine the conditions under which same cohort siblings are suitable for CKMR analysis. A half sibling kinship analysis of single year data set of Atlantic salmon from the River Etne 2013 is performed to check if it is suitable for CKMR.

In the third part, the probability that an individual has a living parent in an age structured population is discussed in detail. For the case where age information for both parent and offspring is unavailable, I derive two useful expressions for the probability of a living parent when mortality is constant, or constant after onset of maturity. With the additional assumption of constant population size, this probability is shown to be 1/2, similar to what is previously proved for constant fecundity.

[1] Ruzzante, D. E., McCracken, G. R., Førland, B., MacMillan, J., Notte, D., Buhariwalla, C., Flemming, J. M., Skaug, H., 2019. Validation of close-kin mark-recapture (CKMR) methods for estimating population abundance. In revision

[2] Skaug, H. J., 2017. The parent–offspring probability when sampling age-structured populations. Theoretical Population Biology 118, 20–26.
URL http://linkinghub.elsevier.com/retrieve/pii/S0040580917300138

Acknowledgement

First of all, thanks to my supervisor, Hans J. Skaug for guidance, interesting discussions, for making me write *en route*, for knowing when to stop, suggestions and ideas, and for encouragement when my own ideas went in other directions. I am also thankful for being given the opportunity to do write my master in the field of statistical ecology, where mathematical statistics and biology merge.

I am very grateful to Daniel Ruzzante for inviting us to join him to work on the Brook trout problem and paper. It has been a good learning experience, and an interesting problem.

Thanks to Kurt Urdal for an update in salmon biology, and some enlightening discussions about the salmon data set and scale reading methods.

Thanks to Alison Harvey and Kevin Glover at the Institute of Marine Research for giving us access to the River Etne salmon data set.

Thanks to Olaug Vetti Kvam and my other colleagues at the Centre for Science Education for giving me room for studying statistics, for being tolerant in the periods when my work schedule have been rather unpredictable during this last year, and for giving me an enjoyable work environment.

Thanks to the Faculty of Mathematics and Natural Sciences for giving me a 50% study leave with 80% pay for this final year of my thesis.

A thank also to Nikolai Østgaard who talked me into taking his course in space physics, which reminded me how fun it is to study new topics. This again motivated me to take further courses in math and statistics.

And finally, thanks to my wonderful family, which have been patient with me when I have occupied the living room table with my computer and papers, and isolated myself from the rest of the world with earbuds. Love you, Oddfrid, Are, Ylva and Varg!

Contents

Chapter 1

Introduction

Close-kin mark-recapture (CKMR) is a method for estimating animal abundance made possible in the last few decades by the increasing availability of DNA based methods (Bravington et al. 2016b). The method builds on the foundations of mark-recapture methods, but extends this by by viewing all sampled individuals as a "mark", and the sampling of a relative of a "recapture" of the marked individuals. Given an appropriate sampling design, the method can also be used to estimate other demographic parameters. Examples of demographic parameters of interest are mortality, population growth rate, or population age structure.

Mark-recapture models includes a wide range of methods for estimating animal abundance and related population parameters, such as survival and population growth rate. The common element is that an animal is observed and "marked" by some mean, and later a portion of the marked animals are "recaptured", one or more times. A mark can be any form of information which makes it possible to identify an animal. Classical marks include methods like physical tags attached to the animals, bird ringing, fish fin clipping, paint marking of animals, and purely observational methods, for example visual identification of individual animals.

As DNA based methods has become more precise and less expensive, genetic marks have been an increasingly important part of the mark-recapture toolbox. DNA based markers makes it possible to mark an animal without attaching a physical mark that may affect the animal. It even allows for indirect mark-recapture without direct observation of the animals, for example via faeces samples or hair samples (Schwartz et al. 2010). CKMR expands the scope of genetic markers in mark-recapture further, by allowing the recaptures to be not only recaptures of the same animal, but also captures of close relatives. CKMR avoids several of the problems with ordinary mark-recapture methods. In ordinary mark-recapture, a fundamental assumption is that the captured animal is unaffected by the marking process, while CKMR is equally applicable with lethal sampling.

Initial development of the CKMR method was done by Skaug (2001), and the method has recently gained wide interest, after both new theoretical developments and practical applications (Bravington et al. 2016a,b). Bravington et al. (2016a) applied CKMR to a southern bluefin tuna population, and demonstrated that the method could be of practical use for abundance estimation. By using half sibling pairs, CKMR can contribute to estimate adult population even in cases where the adult population is extremely hard to sample, as Hillary et al. (2018) show in their white shark study. Bravington et al. (2016b) lists three main statistical challenges when extending mark-recapture to the close-kin case. These are

1. Formulating a demographic model to describe the "recapture" probabilities.
2. Complicated likelihood, a full likelihood would include the while genealogy of the whole population.
3. Quantifying uncertainties in kinship determination.

The goal of the present work is to explore selected aspects of CKMR theory, focusing mostly on the demographic aspects.

During the present work we were approached by Dr. Daniel Ruzzante and Dr. Joanna Mills Flemming at Dalhousie University, Halifax and invited to cooperate on a CKMR validation study. Seven river dwelling brook trout populations were sampled each year from f to 2018 using electrofishing. The study compares the parent-offspring CKMR of the 2014 population with the estimates obtained by ordinary mark-recapture. Chapter 2 presents the development of a parent-offspring age structured CKMR model for this population. Further analysis of the estiator is given in chapter 3. The results of the CKMR analysis are presented in Ruzzante et al. (2019).

1.1 Classical mark-recapture models

Lincoln-Petersen two sample estimator

The simplest classical mark-recapture model is the Lincoln-Petersen, or Petersen method for estimating total population size in a closed population. A population is sampled at two occasions, where at the first occasion n_1 individuals are marked and released, and at the second occasion n_2 individuals are sampled, of which m_2 are marked. The Lincoln-Petersen estimator found by equating the proportion marked individuals in the second sample with the proportion marked individuals in the whole population, $n_1/N = m_2/n_2$ so that

$$\hat{N}_{L-P} = \frac{n_1 n_2}{m_2}. \tag{1.1}$$

If we assume that 1. closed population (N constant) 2. the second sample is a random sample 3. animals do not lose their marks and all marks are recovered, the distribution of $m_2|n_1, n_2$ follows a hypergeometric distribution (A.4), where $m_1 = 0, 1, \ldots min(n_1, n_2)$ (Seber 1982, chapter 3.1.1). The Lincoln-Petersen estimator has a positive bias for small values of m_2. Another problem is that if m_2 is allowed to take the value zero, the expectation and variance of N_{L-P} does not exist. Based on hypergeometric model, Chapman (1951) derived a modified estimator to adjust for the bias,

$$\hat{N}_{Ch} = \frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1. \tag{1.2}$$

An alternative estimator is given by Bailey (1951) (Seber 1982, chapter 3.1.1), using the binomial approximation to the hypergeometric model,

$$\hat{N}_{Ba} = \frac{n_1(n_2+1)}{(m_2+1)}. \tag{1.3}$$

Baileys estimator is nearly unbiased, and has a proportional bias of order $e^{-E(m_2)}$.

Heterogeneity and independence

Dependence between samples can create very biased results in a mark recapture analysis, if not accounted for. There are two causes for dependency bias in mark recapture, 1. local, or list dependence, and 2. individual heterogeneity (Chao et al. 2008). Local dependence arises when there is a causal relationship, and the recapture of an animal is affected by the initial capture, e.g. an animal has a behavioural response and become trap shy or trap happy.

For a random effect model where the P_j and P_k are the capture probabilities of an animal in sample j and k Chao et al. (2008) defines the coefficient of covariation between the samples as

$$CCV(P_j, P_k) = \gamma_{jk} = \frac{COV(P_j, P_k)}{\mu_j \mu_k} = \frac{E(P_j P_k)}{\mu_j \mu_k} - 1. \tag{1.4}$$

Using the approximation $E(n_1 n_2/m_2) \approx E(n_1 n_2)/E(m_2)$ they write

$$N \approx \frac{E(n_1)E(n_2)}{E(m_2)}(1+\gamma_{12}) = E(\hat{N})(1+\gamma_{12}). \tag{1.5}$$

If the only source of heterogeneity is individual variation in capture probability, $\gamma_{12} = CV(P)^2$ (Seber 1982, see also p.86).

We refer to the bias in $E(\hat{N})$ caused by correlated capture probabilities as the *correlation bias*, $E(\hat{N}) - N \approx -\gamma_{12}E(\hat{N})$. This illustrates a general property shared by many capture-recapture models, if the capture probabilities are correlated, this will cause a negative bias in the abundance estimates.

Multiple samples

The extension of mark recapture to multiple samples makes it possible to expand on the information obtained from the mark recapture experiment to other population parameters. The building block for a multiple marking census is the *capture history*, or sighting history, of each observed individual. Inference for population parameters can then be done by from the likelihood of the capture history.

1.2 Close-kin mark-recapture models

The CKMR census is related to ordinary mark-recapture, but there are also several differences which add complexity to the model. In CKMR, the construction of a full likelihood is difficult, because of dependence between pairs from the genealogy and uncertainty in kinship detection, which also is dependent on the genealogy. Instead a pseudo likelihood method is preferred (Skaug 2001; Bravington et al. 2016b), where the pairwise kinship probabilities are used.

1.2.1 Parent-offspring models

Skaug 2001's single sample model

The CKMR approach was first described by Skaug (2001). Assuming random sampling, and using standard population genetics assumptions (random mating, Hardy-Weinberg equilibrium and linkage equilibrium), he derives an estimator for the total population size based on the number of parent offspring pairs, using a pseudo likelihood approach.

The total population is estimated as

$$\hat{N} = \frac{(\rho_m + \rho_f)n(n-1)}{H}, \tag{1.6}$$

where ρ_m and ρ_f is the probability that the father and mother of a randomly selected individual is alive, n is sample size, and H is the number of parent-offspring pairs in the sample. The derivation makes few assumptions of the demography of the population. The final expression does, however, contain two non-trivial parameters which summarize all necessary demographic information, the probabilities that the mother and father of an individual is alive and in the population.

Skaug (2001) suggests using $\rho_m = \rho_f = 1/2$ as an estimate for the probabilities of living parents. In Skaug (2017) he proves that this will hold for an age structured population in the case of constant fecundity. The extension of this result to other cases is the topic of chapter 6.2 below, where a similar result is derived for constant mortality.

Lincoln-Petersen type estimator

Bravington et al. (2016a,b) gives a simple two sample estimator for the adult population size, which is the close-kin version of the Lincoln-Petersen population size estimator. Consider a sample of n_J juveniles and n_A adults from the same population. Each sampled juvenile can be considered a “mark” of its two adult parents. In a population of N_A adults, the probability that a random adult is one of a random juvenile's parent is $2/N_A$. The expected number of parent offspring pairs across all $n_A \cdot n_J$ comparisons is then $2n_An_J/N_A$. Equating this to observed number of parent offspring pairs gives the CKMR version of the Lincoln-Petersen estimator

$$\hat{N}_A = \frac{2\,n_J\,n_A}{H}. \tag{1.7}$$

The basic assumption is that the probability that an adult is caught is independent of the number of offspring. The assumptions needed for this to hold is (Ruzzante et al. 2019), 1. The adults are sampled after, or at the same time that as the juveniles are born and 2. either fecundity is age independent, or all parents have the same probability of being sampled. Further discussion of the assumptions and properties of the moment estimator is given in chapter 3.

Bravington et al. (2016a,b) use the Lincoln-Petersen type estimator mainly as a motivation for the CKMR method, and to show the closeness of the CKMR method to regular mark-recapture. Chapter 2 demonstrates how the moment estimator can be extended to age structured populations, and applied to a brook trout population. A version of the Lincoln-Petersen estimator for parent offspring is also derived and used by Rawding et al. (2014) to estimate the population size in semelparous Pacific salmon species from the post breeding carcasses.

Expected relative reproductive output model

The construction of an estimator for CKMR is simplified by the notion of expected relative reproductive output, introduced by Bravington et al. (2016a,b). The expected relative reproductive output model rests on a simple idea. If we can assume all offspring born at the same time and location have the same probability of getting caught, the probability that a caught offspring is the offspring of a given individual is simply the number of offspring produced by the individual over the total number of juveniles born at the same time and location. The observed heterogeneity in the parent population is handled by conditioning on the observed covariates. The demography of males and females will usually differ, and needs to be considered separately. I will use the convention of considering females and offspring only, the male case has an identical formulation. More specifically, consider a sampled mother i and a sampled offspring j. Let $R_i(\mathbf{y}_j)$, denote the reproductive output, i.e. number of offspring produced by female i at the time and place y_j when j was born, and $B(\mathbf{y}_j)$ the total number of offspring at the same time and place. If we know exactly both the number of offspring produced by a female i at $\mathbf{y}$ and the total number of offspring produced at that location at that time and place, and assume that all offspring born at the time and place $\mathbf{y}_j$ has the same probability of getting sampled, then

$$P(\text{i is j's mother}|\mathbf{y}_j, R_i(\mathbf{y}_j), B(\mathbf{y}_j)) = \frac{R_i(\mathbf{y}_j)}{B(\mathbf{y}_j)}. \tag{1.8}$$

Assumption 1 Given $\mathbf{y}_j$, and i's reproductive output at $\mathbf{y}_j$, the probability that j is sampled is independent of the event that i i sampled, i.e. no heritable heterogeneity in sampling probability or survival given reproductive output. Formally,

$$P(i \text{ is } j\text{'s mother}|\mathbf{y}_j, R_i(\mathbf{y}_j), B(\mathbf{y}_j), i \text{ sampled}, j \text{ sampled}) = P(i \text{ is } j\text{'s mother}|\mathbf{y}_j R_i(\mathbf{y}_j), B(\mathbf{y}_j)). \tag{1.9}$$

I will here let $\mathbf{y}_j$ just mean j's time of birth, but it could just as well mean time and location (Bravington's (x, y)). Both $R_i(\mathbf{y}_j)$ and $B(\mathbf{y}_j)$ are unobserved latent variables. Let $\mathbf{z}_i$ be the observed covariates associated with female i, e.g. size or age. The probability that i is j's mother, given that she is sampled and $\mathbf{z}_i$ can then be found as

$$P(i \text{ is } j\text{'s mother}|\mathbf{y}_j, \mathbf{z}_i, \text{i sampled}) = E_{R_i(\mathbf{y}_j), B(\mathbf{y}_j)|\text{i sampled}, \mathbf{z}_i))} \frac{R_i(\mathbf{y}_j)}{B(\mathbf{y}_j)}. \tag{1.10}$$

As an approximation, we can assume that B is large relative to the individual reproductive output, so that $E(B(\mathbf{y}_j)|R_i(\mathbf{y}_j)) \approx E(B(\mathbf{y}_j))$

$$E_{R_i(\mathbf{y}_j),B(\mathbf{y}_j)|\text{i sampled},\mathbf{z}_i)} \frac{R_i(\mathbf{y}_j)}{B(\mathbf{y}_j)} \approx \frac{E(R_i(\mathbf{y}_j)|\text{i sampled}, \mathbf{z}_i)}{E(B(\mathbf{y}_j))}. \tag{1.11}$$

This simplifies the calculations, and is needed to get the exact form given by Bravington et al. (2016b, eq. 3.5), but not necessary for the model. If B can take very small values, for example if summing over very rare locations, the approximation is invalid.

Assumption 2 Given the observed covariates $\mathbf{z}_i$ for the mother, the fact that i is sampled does alter the posterior distribution of $R_i(\mathbf{y}_j)$,

$$E(R_i(\mathbf{y}_j)|i \text{ sampled}, \mathbf{z}_i) = E(R_i(\mathbf{y}_j)|\mathbf{z}_i). \tag{1.12}$$

This means that there must be no unexplained heterogeneity in i which affect both the probability of i being sampled and i's reproductive output. Combining these,

$$P(i \text{ is } j\text{'s mother}|\mathbf{y}_j, \mathbf{z}_i, i \text{ sampled}, j \text{ sampled}) = P(i \text{ is } j\text{'s mother}|\mathbf{y}_j, \mathbf{z}_i), \tag{1.13}$$

and

$$P(i \text{ is } j\text{'s mother}|\mathbf{y}_j, \mathbf{z}_i) = \frac{E(R_i(\mathbf{y}_j)|\mathbf{z}_i)}{E(B(\mathbf{y}_j))}. \tag{1.14}$$

An interesting property of (1.14) is what does *not* appear in the model. Since we condition on the mother being sampled, the sampling probability does not appear. Likewise, since we know the mother is alive at the time of sampling, the adult mortality does not enter the expression if she is caught after y_j. However, if the potential mother is caught before j is born, her expected output at time y_j must include the survival probability from t_j to y_j. Also, since the survival and sampling probability is assumed to be equal for all offspring born at the same time, it appears in both the nominator and denominator and cancels out.

Note that if survival or sampling probability differs between male and female offspring, the model needs to consider daughters and sons separately, since the model as stated here assumes that sampling probability and survival are equal for all offspring.

1.2.2 Half-siblings

In many cases, obtaining samples from parent-offspring pairs can be difficult. If the length of the adult phase is short compared to the juvenile phase, it will be difficult or impossible to obtain parent-offspring pairs, unless the sampling is done over a long time period. Many organisms might only be possible to sample in certain life phases, for instance only juveniles might be sampled for some species.

We need to consider maternal and paternal siblings separately, as the demography of parenthood in general is very different for the two sexes. In particular, the reproductive variability is usually higher for males than for females. Other demographic parameters like age at maturity and mortality may also differ considerably. For simplicity, I will only consider maternal half siblings (MHS).

For the CKMR application, the relevant probabilistic question is given an individual, what is the probability that a randomly selected individual is the sibling of an individual at hand? Since each pair only needs to counted once, it is only needed to consider either younger or older siblings, in addition to same age siblings. The mothers are unknown, and the probability of sistership must be found by summing over all possible mothers.

Assuming random mating and that the ages of the sampled offspring are known, and ignoring stock structure, Bravington et al. (2016b) gives an equation for the probability that a random individual born at time y_j is the MHS of an individual born at time y_i as

$$\begin{aligned} P(i \text{ and } j \text{ are MHS}) &= \sum_{k \in \mathcal{F}_i} P(i\text{'s mother is } k) P(j\text{'s mother is } k) \qquad (1.15) \\ &= \sum_{k \in \mathcal{F}_i} \frac{R_k(y_i)}{B(y_i)} \frac{R_k(y_j)}{B(y_j)}, \qquad (1.16) \end{aligned}$$

where $R_k(y)$ is the reproductive output of female k at time y, and $\mathcal{F}_i$ is the set of possible mothers for i, and $B(t)$ is the total number of births in the population at time t. This simple expression hides a lot of demography. A general model would have to describe the demographic changes of both the individual females and the total population between the years y_i and y_j.

The reproductive output term here needs some explanation. We assume that the pairs are ordered so that $y_i \leq y_j$ and sums over the possible mothers of i. The expected reproductive output of R_k at time y_i is then simply her expected fecundity F_{x_i}, but the expected reproductive output at time y_j must be discounted by the probability of surviving from y_i to y_j.

Reproductive output is a random variable. Since we are summing over individuals, the reproductive outputs are not independent. For same year siblings, the probabilities are dependent on both the value of the random variable and its variance. Similarly, if the reproductive output is dependent between years, the probabilities are dependent on the covariances. Bravington et al. (2016b) suggest that the problem of variance in the same year can be avoided by only considering sibling born in different years, and notes that individual heterogeneities in fecundity consistent over time will need to be modelled to avoid bias in the probabilities.

Another reason why different cohort siblings are of particular interest, is that they constitute marking of a parent at two points in time, and contains information about the mortality of the parent between markings. In the base case where fecundity and mortality is constant,

$$P(i \text{ and } j \text{ are MHS}) = \frac{S(y_i, y_j)}{N_{♀}(y_j)}, \qquad (1.17)$$

where $S(y_i, y_j)$ is the survival probability from y_i to y_j, and $N_{♀}(y_j)$ is the population size at the time j (the younger sibling) is born (Bravington et al. 2016b, (3.1)).

Half sibling CKMR is discussed further in chapter 4 and 5.

1.3 Demographic modelling

Fecundity, survival and other demographic parameters, such as population growth rate, plays an essential part in CKMR modelling, both on the individual level and on the population level. If we consider the expected relative reproductive output model (1.14), on the individual level we need to estimate the expected number of offspring for the parent at the time of birth of the offspring. On the population level, we need to estimate the total number of births at the same point in time, which is a function of both the population size and demography at the time.

1.3.1 Parent- vs offspring-centric approach

Skaug (2017) makes the distinction between an offspring-centric and a parent-centric approach to derive the parent-offspring probabilities. The expected relative reproductive output model is offspring-centric, since it starts with the offspring and ask for the probability that a given individual is the mother. As an alternative, he consider the parent-centric approach, where the question is framed the other way around. From this, given the daughters age a_j, he derive the decomposition

$$P(j\text{'s mother is } i|a_j) = P(j\text{'s mother is alive}|a_j)P(j\text{'s mother is } i|, a_j, j\text{'s mother is alive}). \tag{1.18}$$

The simple form of (1.14) relies on being able to determine the ages of both the offspring and of the parents. If the age of the parent or offspring is unknown, we need to use demographic modelling to describe the age composition and age specific fecundities in the population. Whether the mother-centric or daughter-centric approach is most convenient may depend on the specific case. The mother-centric approach can also be used when the assumptions leading to (1.14) does not hold (Skaug 2017). Chapter 6 is devoted to the probability of having a living mother in continuous time stable age models, and derives several results for the case where age information is unavailable for both parent and offspring.

Central to demographic theory is stable age theory, which relates the individual level age specific fecundities and survival probabilities to the age distribution and growth rate of the whole population. Chapter 6.1 give a review of the basic results from stable age theory for the continuous case.

1.3.2 Demographic parameters and notation

Fecundity and expected reproductive output

The expected relative reproductive output model above is specified in terms of the notion of $R_i(y)$, the number of offspring produced by the mother i at the time y_j. Here, I will use the term F, or fecundity at age $F(a)$, to denote the number of offspring produced by an individual aged a. The relationship can be expressed as

$$R_i(y_j) = \mathbb{I}(i \text{ survives to } y)F(i\text{'s age at } y), \tag{1.19}$$

Table 1.1: Main notation used in discrete time models.

N	Population size
n	Sample size
a	Age
$p(a)$	Relative frequency by age
F	Fecundity
S	Survival
H	Number of matched kin pairs
t	Time of capture
y	Time of birth
$_A$	Adults
$_J$	Juveniles
B	Total number of births in the population

where $\mathbb{I}(i \text{ survives to } y)$ is an indicator variable for the event that i survives. Both $\mathbb{I}(i \text{ survives to } y)$ and $F(i\text{'s age at } y)$ are random variables. Taking the expectation of (1.19) gives

$$E(R_i(y_j)) = P(i \text{ survives to } y)\bar{F}(i\text{'s age at } y). \tag{1.20}$$

The covariates $\mathbf{z}$ in (1.14) will not appear directly in the formulas in later chapters. The parent-offspring CKMR model is for an age structured model, and instead age and birth year will be used directly. This correspond to assuming that the observed covariates (size, scale samples, etc) are sufficient to determine age with certainty.

Chapter 2

Brook trout CKMR estimator

This chapter gives a simple extension of the Lincoln-Petersen type estimator for an age structured population. The model is developed for a design study by Ruzzante et al. (2019) on stream dwelling brook trout (*Salvelinus fontinalis*) populations in Novia Scotia, comparing the CKMR census with ordinary mark-recapture census.

The study sites and populations are described in Ruzzante et al. (2016, 2019). The example data used below are from the RCU population. The model is developed for specifically for the Brook trout population, but should be equally applicable to other closed, age structured populations.

This chapter gives a basic derivation of the formulas using the expected reproductive output model from Bravington et al. (2016b). The next chapter further explores the properties of the estimator and gives an alternative derivation and a more general version of the estimator. Only the expectation of the fecundity or reproductive output is used in this chapter, and I will therefore use a simplified version of the notation in section 1.3.2, and let F refer to the expectation of the fecundity (see table 2.2).

2.1 Problem description

We have a data set consisting of juveniles belonging to a single cohort observed in a single year. These are are matched with parents observed the year before the juveniles are born, the same year, and in the succeeding years. Age of all captured individuals are known. We also have independently obtained estimates of yearly survival and fecundity for each age class. Sex of the captured individuals are unknown, and will have to be treated as a latent variable.

Using this information, we shall estimate the adult population size in the birth year of the juvenile cohort.

Table 2.1: Number of sampled fish by year and age group for the RCU population. All samples in green cells are potential parents, the blue cell are samples of the juvenile cohort born in 2015.

Age	2014	2015	2016	2017
1+	46	23	15	93
2+	13	49	64	51
3+	0	8	10	2

Table 2.2: Notation for the brook trout model.

t_0	Juvenile cohort's birth year
a_0	Parents age at time t_0
$N(a,t)$	Population size of year class a at time t
$n(a,t)$	Number of sampled individuals
$F(a)$	Relative fecundity age a
$S(a)$	Survival from age class a to $a+1$
S_{t_0}	Survival from sampling to t_0
n_J	Number of sampled juveniles
$P_H(a,t)$	Probability that an individual parent to a given juvenile
$H(a,t)$	Number of matched parents
$R_{t_0}(a,t)$	Expected reproductive output of an individual in t_0
N_A	Total adult population size in year t_0

2.2 Expected reproductive output model

If we consider each cell (age-time combination) of captured potential parents separately, we can then find the probability that an individual i aged a in the sample is the mother of a sampled juvenile from (1.14) as

$$P(i \text{ is } j\text{'s mother}) = \frac{i\text{'s expected reproductive output in } t_0}{\text{Total reproductive output in } t_0} = \frac{R(a,t_0)}{B(t_0)}, \tag{2.1}$$

where t_0 is the birth year of the juvenile cohort. Summing over both parents,

$$P(i \text{ is } j\text{'s parent}) = P(i \text{ is } j\text{'s mother}) + P(i \text{ is } j\text{'s father}). \tag{2.2}$$

If the life history parameters differ for males and females, these probabilities must be considered separately. If the a priori probability that a captured adult is female is c,

$$P(i \text{ is } j\text{'s parent}) = cP(i \text{ is } j\text{'s father}|\text{male}) + (1-c)P(i \text{ is } j\text{'s mother}|\text{female}). \tag{2.3}$$

2.3 Total reproductive output

Let $N(a,t)$ denote the number of individuals aged a at time t, and t_0 be the birth year of the juvenile cohort. Assume fecundity to be dependent on age only, and let $F(a)$ be the fecundity of an individual in age class a. Let $p(a) = N(a,t_0)/\sum_{x>0+} N(x,t_0)$ be the proportion of the adult population in age class a, and $N_A = \sum_{x>0+} N(x, t_0)$ be the total adult population at time t_0. We can find the expected total reproductive output in year t_0 by summing over the expected reproductive output for the females in each age class. Allowing for unequal fecundity in males and females, but assuming an even sex ratio,

$$B(t_0) = \sum_{x>0+} F_{\female}(x)N(x, t_0)/2 = N_A \sum_{x>0+} p(x)F_{\female}(x) \triangleq N_A\bar{F}_{\female}/2. \tag{2.4}$$

Total number of offspring produced by the fathers in the population is clearly the same as the number produced by the mothers, $B(t_0) = N_A\bar{F}_{\female}/2 = N_A\bar{F}_{\male}/2$, and therefore $\bar{F}_{\female} = \bar{F}_{\female}$, if the sex ratio is even. However, since the expression (2.1) only contains relative reproductive output, it is still useful to consider $\bar{F}$ separately for each sex, as this allows us to use different relative measures of fecundity for males and females.

2.4 Parent's reproductive output

As noted above, we only need relative fecundities, and we can therefore use $F(a_0)$, fecundity of the parent cohort in question as a measure of reproductive output of the parents. For the potential parents observed the year before the birth, the fecundity must be discounted by the probability that the individual survives to reproduce, i.e. $S(a_0 - 1)F(a_0)$. For the potential parents belonging to the cohort aged a_0 in t_0, and observed in t_0 or after,

$$P(i \text{ is } j\text{'s parent}) = 2\frac{F(a_0)}{\sum_{x>0+} F(x)N(x, t_0)} = \frac{2F(a_0)}{N_A\bar{F}}, \tag{2.5}$$

and for the parents observed the year before birth of the cohort,

$$P(i \text{ is } j\text{'s parent}) = 2\frac{S_{a_0-1}F(a_0)}{\sum_{x>0+} F(x)N(x, t_0)} = \frac{2S_{a_0-1}F(a_0)}{N_A\bar{F}}. \tag{2.6}$$

Unequal fecundity at age in males and females

Ref. (2.3), and assuming even sex ratio,

$$P(i \text{ is } j\text{'s parent}) = \frac{F_{\female}(a_0)}{N_A\bar{F}_{\female}} + \frac{F_{\male}(a_0)}{N_A\bar{F}_{\male}}. \tag{2.7}$$

$\bar{F}_{\female}$ and $\bar{F}_{\male}$ is the same number if measured in number of offspring. However, it is usually easier to use a proxy that scales with the number of offspring, e.g. body weight, which may differ between the sexes. Note that the assumption of equal sex ratio at all age classes also implies equal mortality for males and females.

2.5 Age structure

To calculate the total reproductive output of the population at time t_0, we need to estimate the reproductive output of each of the age classes. To do this, we need to know the age specific fecundity, and the proportion of the population in each age class ($p(x)$, (2.4)). For a larger data set with more parent offspring pairs, $N(a_0, t_0)$, $a_0 = (1+, 2+, 3+)$, could be estimated as separate parameters using maximum likelihood and the probabilities above. Using only the data at hand, there are two possibilities:

1. Assume constant population size N_A and varying sample selectivity
2. Assume that the yearly samples give an estimate of the population age structure, and that recruitment is variable

1 Constant population size

Assuming constant population size and mortalities, and three reproductive age classes, we can express the later adult age classes relative to the first adult age class, $N_1 = N(1, a_0)$:

$$\begin{aligned} N_2 &= S_1 N_1, \\ N_3 &= S_1 S_2 N_1, \end{aligned} \tag{2.8}$$

and total adult population size as

$$N_A = N_1 + N_2 + N_3 = N_1(1 + S_1 + S_1 S_2). \tag{2.9}$$

The proportion in each age class is then

$$\begin{aligned} p_1 &= N_1/N_1(1 + S_1 + S_1 S_2) = 1/(1 + S_1 + S_1 S_2), \\ p_2 &= S_1/(1 + S_1 + S_1 S_2), \\ p_3 &= S_1 S_2/(1 + S_1 + S_1 S_2). \end{aligned} \tag{2.10}$$

The total reproductive output at t_0 can then be found as

$$\begin{aligned} B(t_0) &= N_A \frac{F_1 + S_1 F_2 + S_1 S_2 F_3}{1 + S_1 + S_1 S_2} \\ &= N_A \bar{F}_c, \end{aligned} \tag{2.11}$$

where $\bar{F}_c$ is the average fecundity we find from assuming stable population structure.

2 Variable recruitment

An alternative to assuming stable population size and structure is to assume that the sampling gives a correct picture of the relative abundance of the age classes.

Assuming this we can use the capture data as estimate of year class abundance, and normalise by total capture,

$$p(x) = \frac{n(x, t_0)}{\sum_{x>0+} n(x, t_0)}, \tag{2.12}$$

and

$$B(t_0) = \sum_{x>0+} F(x) N_A \frac{n(x,t_0)}{\sum_{x>0+} n(x,t_0)} = N_A \bar{F}_v, \tag{2.13}$$

where N_A now refers to the total adult population size at time t_0 only, and $\bar{F}_v$ is the average fecundity at time t_0 if we assume that the sample is representative for the population structure.

2.6 Moment estimator for N_A

To simplify notation and calculation, let $S_{t_0}(a,t)$ be the probability that an individual aged a at time t will survive to the year of the birth of the juvenile cohort, t_0,

$$S_{t_0}(a,t) = \begin{cases} S(a) & t = t_0 - 1 \\ 1 & t \geq t_0 \end{cases}. \tag{2.14}$$

Using the above, let $P_H(a,t)$ be the probability that an adult in age class a sampled at time t is the parent of one of the juveniles,

$$P_H(a,t) = 2\frac{S_{t_0}(a,t)F(a-(t-t_0))}{N_A \bar{F}}. \tag{2.15}$$

Let $R_{t_0}(a,t)$ be the expected reproductive output in year t_0 of an individual sampled as age a in year t,

$$R_{t_0}(a,t) = S_{t_0}(a,t)F(a-(t-t_0)). \tag{2.16}$$

The expected number of parent offspring pairs in any age-time cell is then

$$\begin{aligned} E(H(a,t)) &= n_J n(a,t) P_H(a,t) \qquad (2.17) \\ &= n_J n(a,t) \frac{2R_{t_0}(a,t)}{N_A \bar{F}}. \qquad (2.18) \end{aligned}$$

A moment estimator for N_A can be found by summing over all cells, using the expected reproductive output as a weight, and setting the observed number of parent offspring pairs H equal to the expected,

$$\hat{N}_A = \frac{2n_J}{H_{tot}\bar{F}} \sum_{a=1}^{3} \sum_{t=t_0-1}^{t_0+2} R_{t_0}(a,t) n(a,t). \tag{2.19}$$

Here $\bar{F}$ may refer to $\bar{F}_v$ or $\bar{F}_s$ above, depending on the assumed age structure. Alternatively, summing diagonally along each parent cohort,

$$\hat{N}_A = \frac{2n_J}{H_{tot}} \sum_{a=1}^{3} \frac{F(a)}{\bar{F}} \sum_{t=t_0-1}^{t_0+2} S_{t_0}(a+(t-t_0),t) n(a+(t-t_0),t). \tag{2.20}$$

Allowing for unequal fecundity at age in males and females,

$$\hat{N}_A = \frac{n_J}{H_{tot}} \sum_{a=1}^{3} \left(\frac{F_{♀}(a)}{\bar{F}_{♀}} + \frac{F_{♂}(a)}{\bar{F}_{♂}} \right) \sum_{t=t_0-1}^{t_0+2} S_{t_0}(a+(t-t_0),t)n(a+(t-t_0),t). \tag{2.21}$$

If we only observe the potential parents after birth of the cohort, and fecundity is equal for all age classes, $R_{t_0}(a,t) = \bar{F}$, and the formula reduces to the simpler formula (1.7),

$$\hat{N}_A = \frac{2n_J n_A}{H_{tot}}. \tag{2.22}$$

However, applying this directly to the age structured population will give a biased result, and we need to take into account that parents in the different age classes have different probabilities being “marked” by a juvenile. Note that n_J and the other captures enters the equations as the number of possible matched pairs. There is thus no need to adjust for the same parent having matches with more than one offspring.

Bias correction

Like the ordinary Lincoln-Petersen estimator, this estimator, $\hat{N}$, has a positive bias. This is caused by H being in the denominator of the expression, and might be quite large for small H. A nearly unbiased version of the estimator can be achieved by replacing H with $H+1$ in the denominator,

$$\hat{N}_{A_C} = \frac{n_J}{(H_{tot}+1)} \sum_{a=1}^{3} \left(\frac{F_{♀}(a)}{\bar{F}_{♀}} + \frac{F_{♂}(a)}{\bar{F}_{♂}} \right) \sum_{t=t_0-1}^{t_0+2} S_{t_0}(a+(t-t_0),t)n(a+(t-t_0),t). \tag{2.23}$$

Further discussion of the bias correction and derivation of this result is given in chapter 3.2.

2.7 Example

This example use the catch data in table 2.1, from Ruzzante et al. (2019). The sampled juveniles are the 1+ age class born in 2015 and sampled in 2017.

Her P_s is the proportion in each age class estimated assuming a stable age distribution and constant population, and P_v is the proportion in each age class from assuming that the catch data for year 2015 in table 2.1 is a representative sample of the age structure in this year. Using the values in the table above, $\bar{F}_s = 2.34$ and $\bar{F}_v = 4.11$. The difference is quite large, but this is to be expected in this case. If we assume stable age structure, the less fecund class 1+ is the largest age class, and if we use the catches as an estimate of age class, most individuals belong to the fully mature age classes. The sum expression in (2.19) can be calculated by multiplying and summing the entries of the table of catches (table 2.1) and the table of reproductive output (table 2.3). Assuming stable age distribution and a total of for example $H_{tot} = 7$ matched pairs,

Table 2.3: Reproductive output weights for the 2015 population, R_{t_0}

	2014	2015	2016	2017
1+	$S_1 F_2$	F_1	0	0
2+	$S_2 F_3$	F_2	F_1	0
3+	0	F_3	F_2	F_1

Table 2.4: Expected reproductive output in 2015, R_{t_0} (survival to 2015 × fecundity in 2015), relative to a 1+ in year 2015.

	2014	2015	2016	2017
1+	2.08	1.00	0.00	0.00
2+	0.85	4.95	1.00	0.00
3+	0.00	7.73	4.95	1.00

Table 2.5: Demographic parameters for the RCU population used in the analysis. Subscripts v and s refers to model with variable recruitment or stable age structure.

Age class	Fecundity	Rel. fec	S	P_s	P_v	$P_s N_s$	$P_v N_v$
1+	18.395	1	0.42	0.68	0.29	4250	1026
2+	91.08	4.95	0.11	0.29	0.61	1785	2186
3+	142.14	7.73	0	0.03	0.1	196	357

the total population size can be estimated with the bias corrected estimator as

$$\hat{N}_{A_C(s)} = \frac{2 \cdot 93 \cdot 549.6559}{(7+1) \cdot 2.343853} = 5452.347, \tag{2.24}$$

or assuming variable recruitment and representative sampling,

$$\hat{N}_{A_C(v)} = \frac{2 \cdot 93 \cdot 549.6559}{(7+1) \cdot 4.092909} = 3122.351. \tag{2.25}$$

We can find a confidence interval for $\hat{N}_{A_C(v)}$ by using that H is approximately Poisson distributed (see section 3.1). From the *poisson.test* function in the R stats package (R Core Team 2018), a 95% confidence interval for EH when $H = 7$ is $(2.814, 14.423)$. Plugging the end points of the confidence interval for H into the estimator for N_{A_C}, we obtain $(1619.61, 6548.62)$ as a 95% confidence interval for $\hat{N}_{A_C(v)}$.

2.8 Maximum likelihood estimator

As an alternative to the simple moment estimator above, the observed catches in each cell can be used to find a maximum likelihood estimator, which in general will have nicer statistical properties.

The number of matched parents in a cell, $H(a, t)$, can be viewed as a binomial experiment with $n(a,t)n_J$ trials and probabilities $P_H(a,t)$ above,

$$H(a,t) \sim Binomial(n(a,t)n_J, P_H(a,t;N_a)), \tag{2.26}$$

To find the maximum likelihood estimator of N_A, maximise the joint log-likelihood of the observations w.r.t. N_A.

$$\mathcal{L}(N_A) = \sum_{a,t} \log(\text{pbinom}(H(a,t), n(a,t)n_J, P_H(a,t;N_A)). \tag{2.27}$$

Here $pbinom(x,N,p)$ is defined as in R, i.e. the probability of drawing x successes from N trials with probability p.

Chapter 3

Properties of the CKMR estimator

In the previous chapter, only the expectation of the reproductive output was used to derive the estimator. In this chapter, I will regard the fecundity as a discrete random variable which describe the offspring count at some interval after the birth. The theoretical properties of the estimator is examined, both generally and in a small population setting. Finally, the chapter contains a discussion the estimator's relation to the regular Lincoln-Petersen estimator and the pairwise pseudo likelihood approach.

3.1 Distribution of parent-offspring pairs (H)

In ordinary closed population mark recapture, a hypergeometric model is often used to describe the recapture probabilities. If the population is large, so that the probability of recapturing the same individual is negligible, a binomial approximation is often used. In CKMR, the distribution of H is not hypergeometric, since the number of marks in the population is also a random variable. An approximate distribution of H can be found from large sample approximations. In a large population setting with sparse sampling, the probability that a random juvenile-adult pair is a parent-offspring pair is a Bernoulli variable with low probability, and also approximately independent (Bravington et al. 2016a, Supp. note 3). The number of observed parent offspring pairs is therefore approximately Poisson distributed, by the law of rare events. The Poisson distribution is a one parameter distribution, with variance equal to the expectation, a property which is used several times in the discussion below. The true distribution of H might deviate from this in both directions, however as discussed below, the deviations are expected to be small.

The direction of marking in CKMR is arbitrary. To understand the variance of the number of pairs sampled, it is most convenient to view it from the direction that the adults mark the juveniles. If M is the number of "marked" juveniles in the juvenile population, the number of "recaptures", H, can be assumed to follow a hypergeometric distribution, conditional on the number of marks, exactly analogous to ordinary mark recapture. By viewing the mark-recapture model this way, we can formulate the model from a population perspective, rather than the pairwise approach used above.

Let $M = \sum_i F_i$ be the number of marked offspring in the population, i.e. number of offspring from one of the parents in the adult sample, and let $\bar{F}$ and $\bar{F}_S$ be the average number of offspring produced by the parents in the population and sample, $\bar{F}_S = E\sum_i F_i/n_A$ and $\bar{F} = N_J/N_A$.

If the sampling probability of all offspring is equal, the number of offspring caught can be described by the hypergeometric distribution, conditional on the number of marked offspring,

$$EH|M = n_J \frac{M}{N_J} \tag{3.1}$$

$$EH = \frac{n_J}{N_J} EM = \frac{n_J}{N_J} E \sum_i F_i = \frac{n_J\, n_A}{N_J} \bar{F}_S \tag{3.2}$$

$$= \frac{n_J n_A}{N_A} \frac{\bar{F}_S}{\bar{F}}. \tag{3.3}$$

Rearranging this and replacing expectations with actual values, gives the single sex version of (2.19) in a different notation,

$$\hat{N} = \frac{n_A n_J}{H} \frac{\bar{F}_S}{\bar{F}}. \tag{3.4}$$

Two sex estimator

The above considers males and females separately. The two sex estimator where the sex of the sampled adults is unknown in section 2.6, can be derived by adding the parent-offspring pairs from both parents. Let c denote the sex ratio, here defined as the probability that a randomly selected individual is female, and assume equal sampling probability for males and females. Then,

$$EH = \frac{n_J n_A\, c}{N_A c} \frac{\bar{F}_{S♀}}{\bar{F}_♀} + \frac{n_J n_A (1-c)}{N_A (1-c)} \frac{\bar{F}_{S♂}}{\bar{F}_♂} \tag{3.5}$$

$$= \frac{n_J n_A}{N_A} \left(\frac{\bar{F}_{S♀}}{\bar{F}_♀} + \frac{\bar{F}_{S♂}}{\bar{F}_♂} \right). \tag{3.6}$$

Replacing expected with observed and rearranging again gives

$$\hat{N}_A = \frac{n_J n_A}{H} \left(\frac{\bar{F}_{S♀}}{\bar{F}_♀} + \frac{\bar{F}_{S♂}}{\bar{F}_♂} \right), \tag{3.7}$$

as expected.

Variance

The variance of H can be split into two components, $VAR(H) = VAR_M(E(H|M)) + E_M(VAR(H|M))$, where the first term describes the variance caused by the variation in the number of marks, i.e. variation in reproduction in the caught individuals, and the second term describe the variance in "recaptures", and can be described by a hypergeometric model. If F_S is the fecundity of a parent conditionally on being sampled, $VAR(M) = VAR(\sum F_i) = n_A VAR(F_S)$, assuming the fecundities are independent. The total variance of H is

$$VAR(H) = VAR_M(E(H|M)) + E_M(VAR(H|M)). \tag{3.8}$$

From the expectation of the hypergeometric distribution (A.5), $EH|M = Mn_J/(N_A\bar{F})$, and

$$\begin{aligned} VAR_M(E(H|M)) &= \frac{n_J^2 VAR(M)}{(N_A\bar{F})^2} && (3.9)\\ &= \frac{n_J^2\, n_A\, VAR(F_S)}{(N_A\bar{F})^2} && (3.10)\\ &= EH\frac{n_J}{N_A\bar{F}}\frac{VAR(F_S)}{\bar{F}_S}. && (3.11)\\ & && (3.12) \end{aligned}$$

If we assume n_J/N_J is small, we can ignore the probabilities of recaptures of juveniles, and can approximate $VAR(H|M)$ using the binomial distribution. Hence,

$$\begin{aligned} E_M(VAR(H|M)) &\approx E_M\left(n_J\frac{M}{N_A\bar{F}}\left(1-\frac{M}{N_A\bar{F}}\right)\right) && (3.13)\\ &= EH - \frac{n_J}{(N_A\bar{F})^2}E(M^2) && (3.14)\\ &= EH - \frac{n_J}{(N_A\bar{F})^2}(VAR(M)+(EM)^2) && (3.15)\\ &= EH - \frac{n_J n_A}{(N_A\bar{F})^2}VAR(F_S) - \frac{n_J(n_A\bar{F}_S)^2}{(N_A\bar{F})^2} && (3.16)\\ &= EH - EH\frac{VAR(F_S)}{\bar{F}_S\bar{F}N_A} - \frac{(EH)^2}{n_J}. && (3.17) \end{aligned}$$

In total,

$$\begin{aligned} VAR(H) &= EH\left(1-\frac{EH}{n_J}+\frac{n_J-1}{N_A\bar{F}}\frac{VAR(F_S)}{\bar{F}_S}\right) && (3.18)\\ &= EH\left(1-\frac{n_A\,\bar{F}_S}{N_J}+\frac{n_J-1}{N_J}\frac{VAR(F_S)}{\bar{F}_S}\right). && (3.19) \end{aligned}$$

If the number of offspring produced by the captured parents is completely known, $VAR(F_S) = 0$, and the CKMR model is equal to the ordinary mark recapture model. In cases with high fecundity, e.g. fish, it is reasonable to assume that the number of marked offspring in the population is much higher than the number of sampled offspring, $n_A\bar{F}_S \gg n_J$. If F_S is Poisson, $VAR(F_S)/\bar{F}_S = 1$, and $VAR(H)$ is slightly less than the Poisson variance. For sparse sampling n_J/N_J is small, and heterogeneity in fecundity will have little effect on the variation of H, unless F_S is very overdispersed.

The two descriptions of the distribution of H are not contradictory, as the hypergeometric distribution is approximately equal to the Poisson distribution if the population is large and sampling is sparse.

Mortality

There is some ambiguity as to what constitutes the original “marked” population here, i.e. at what time the fecundity F here refers to. The main assumption of the model is that all offspring have the same probability of being sampled. If there is mortality which affects the whole clutch, this should to be included in the fecundity to avoid heterogeneity in capture probability. Mortality which affects all offspring equally will cancel out from (3.3), and will not affect the estimate.

If the mortality is random, the surviving population can be regarded as a single random sample of the original offspring population. The juvenile sample are assumed to be a random sample from this population, and is therefore a random subsample from a random sample from the original juvenile population (Seber 1982, ch. 3.2.1).

3.2 Bias correction of $\hat{N}$

The ordinary Lincoln-Petersen estimator asymptotically unbiased, but is known to be biased for small sample sizes. It follows from Jensen's inequality that all estimators of this type will likely have a positive bias,

$$E\hat{N} = E(2n_JC/H) = 2n_JC\,E(H^{-1}) = N\,E(H)\,E(H^{-1}) \geq N. \tag{3.20}$$

Using the approach in Bailey (1951), we can find the approximate bias of $\hat{N}$ by expanding the estimator around $\mu = EH$ and setting $n_JC = N\mu$,

$$\begin{aligned}\hat{N}(H) &\approx Cn_J\left(\frac{1}{\mu} + \frac{1}{\mu^2}(H-\mu) + \frac{1}{\mu^3}(H-\mu)^2 + O(H^{-3})\right) &(3.21)\\ &= N\left(1 - \frac{H-\mu}{\mu} + \frac{(H-\mu)^2}{\mu} + O(H^{-2})\right). &(3.22)\end{aligned}$$

Taking expectation on both sides and ignoring the higher order terms,

$$E(\hat{N}) \approx N\left(1 + \frac{VAR(H)}{\mu^2}\right). \tag{3.23}$$

Ordinary Lincoln-Petersen mark-recapture is well described by the hypergeometric model (see appendix (A.4)). Bailey uses a binomial approximation to the hypergeometric model, and suggests the modified estimator (1.3). Similarly, if we can approximate the distribution of H by a Poisson distribution, $VAR(H) = E(H)$, and (3.23) gives

$$E(\hat{N}_A) \approx N\left(1 + \frac{1}{EH}\right). \tag{3.24}$$

To adjust for this bias, a possible corrected version of the estimator is

$$\hat{N}_{A_C} = \frac{n_JC}{H+1}. \tag{3.25}$$

If H has a Poisson distribution, with $EH = \mu$, we can find the expectation of $\hat{N}_{A_C}$ exactly,

$$E\hat{N}_{A_C} = N\mu \sum_{h=0}^{\infty} \frac{e^{-\mu}\mu^h}{(h+1)h!} \tag{3.26}$$

$$= N \sum_{h=1}^{\infty} \frac{e^{-\mu}\mu^h}{h!} \tag{3.27}$$

$$= N(1 - e^{-\mu}), \tag{3.28}$$

which means $\hat{N}_{A_C}$ is nearly unbiased even for small EH. It is worth noting that the bias is equal to $P(H = 0)$, so if we condition on $H > 0$ the estimator is unbiased. A variant of bias correction above was derived by Chapman (1952) (see also p. 139 Seber (1982)).

3.3 Variance and uncertainty of $\hat{N}$

If we use the Poisson approximation to describe, H, $CV(\hat{N}) \approx 1/\sqrt{EH}$ (Bravington et al. 2016b). This approximation follows from the delta method. If $\hat{N} = Cn_J/H$,

$$VAR(\hat{N}) = VAR(\frac{Cn_J}{H}) \approx (Cn_J)^2 \frac{VAR(H)}{(EH)^4} \tag{3.29}$$

if $H \sim Poisson$, then $VAR(\hat{N}) \approx (Cn_J)^2/(EH)^3 = \hat{N}^2/EH$. If H is allowed to be zero, the expectation and variance of $\hat{N}$ does not exist, we will therefore have to condition on $H > 0$ for the simulations below, where the runs with $H = 0$ is discarded from the calculations. By the same argument, the delta method approximation of the variance of N_{A_C} is

$$VAR(\hat{N}_{A_C}) = VAR(\frac{Cn_J}{H+1}) \approx (Cn_J)^2 \frac{VAR(H)}{(EH+1)^4}. \tag{3.30}$$

If H is Poisson,

$$VAR(\hat{N}_{A_C}) \approx N^2 \frac{(EH)^3}{(EH+1)^4}. \tag{3.31}$$

Since EH is unknown, $VAR(\hat{N}_{A_C})$ will have to be estimated by using the plug-in estimator obtained by replacing EH by the observed value of H. If EH is low, the distribution is very skewed, and the delta method estimates of the variance will be biased. An alternative approximation of the variance of $\hat{N}$ can be found by the formula given by Chapman (1951), see also Seber (1982, chapter 3.1.1 and 4.1.2).

$$VAR(\hat{N}_{A_C}) \approx N^2 \left(\frac{1}{EH} + \frac{2}{(EH)^2} + \frac{6}{(EH)^3} \right). \tag{3.32}$$

However, if EH is low, the estimator obtained by replacing EH with the observed H will have a random value in the denominator, and therefore have a positive bias, as discussed above.

Confidence intervals

If H is reasonably high, confidence intervals for $\hat{N}$ can be found by using the normal approximation and the delta method variance. If H is low, the distribution of N will be heavily right tailed and non-normal, and the delta method variance will underestimate the variance. The numerical simulations in section 3.5 give examples of this, see table 3.4 and fig. 3.1 for an example using the population parameters for the RCU population discussed earlier.

An alternative approach for low H is to use the Poisson approximation directly. By using that H is approximately Poisson, we can find a confidence interval for EH from the Poisson distribution, and plug the endpoints of the Poisson confidence interval into $\hat{N}(H)$ to obtain a confidence interval for $\hat{N}$. An example for the RCU population is given in section 2.7. For the Chapman estimator, Seber (1982, chap. 3.1.1) suggests using the Poisson approximation when then number of recaptures is less than 50, and the normal approximation when the number is above 50.

3.4 Sensitivity to fecundity

As discussed earlier, the variation in fecundity only matters if this variation is correlated with the probability of recapture. In the model above where we assume the catches are representative of the age structure, the fecundity cancels out from the expression if we only consider the catches in year t_0 (2015):

$$\hat{N} = \frac{2n_J}{H\bar{F}} \sum_{a=1}^{3} F(a) n_a(a) = \frac{2n_J n_A}{H}. \tag{3.33}$$

For the other years, fecundity estimates used only to adjust for the change in the distribution of reproductive output in t_0 of the recaptured parents caused by age specific mortality. If we use the model that assumes stable age distribution above, the fecundity estimates have a much higher impact on the estimates, because then the estimated fecundities are also used to correct for a selectivity of catches with respect to the age classes. More generally, the degree to which $\hat{N}$ depends on the fecundity is dependent on the relationship between sampling and fecundity. Let s_i be the sampling probability, for an individual in age class i, and p_i the proportion in age class i. $\bar{F}_s = \sum_i s_i p_i F_i / \sum s_i p_i = \sum_i s_i p_i F_i / \bar{s}$ and $\bar{F} = \sum p_i F_i$

$$\frac{\partial \hat{N}}{\partial F_i} = \frac{2n_A n_J}{H} \frac{p_i}{\bar{F}} \left(\frac{s_i}{\bar{s}} - \frac{\bar{F}_S}{\bar{F}} \right) \tag{3.34}$$

$$= \hat{N} p_i \left(\frac{s_i}{\bar{s}\bar{F}_S} - \frac{1}{\bar{F}} \right). \tag{3.35}$$

If sampling selectivity is low, $(s_i/\bar{s} - \bar{F}_S/\bar{F})$ is low, and $\hat{N}$ is more robust to errors in fecundity estimates.

3.5 Numerical simulations

To examine the bias of $\hat{N}_A$ and $\hat{N}_{A_C}$, and to check the validity of $1/\sqrt{H}$ as an estimate of $CV(\hat{N})$, we can use Monte Carlo simulations. The approach here is a parametric bootstrap procedure, using the estimated demographic parameters from the brook trout populations.

If all adults are sampled in or after the birth year of the juvenile cohort, then

$$\hat{N}_A = \frac{n_J}{H_{tot}} \sum_{a=1}^{3} \left(\frac{F_{♀}(a)}{\bar{F}_{♀}} + \frac{F_{♂}(a)}{\bar{F}_{♂}} \right) n_{tot}(a), \tag{3.36}$$

where $n_{tot}(a) = \sum_t n(a + (t - t_0), t)$ is the sum of all catches of adults from the cohort aged a in the year t_0.

Mortality is not explicitly modelled in the simulation, instead the catches for each cohort were summed, and the cohort born before sampling was weighted by the mortality. The n_i values in table 3.1 thus correspond to the inner summation in (2.20).

Procedure for simulations

1. Given N_A, and age distribution $\mathbf{p} = (p_1, p_2, p_2)$, draw $N_A/2$ females and $N_A/2$ males with with $age \sim multinomial(1, \mathbf{p})$.
2. Fecundity of each individual is set as a function of age, with $F_{♀}(a) = F_{♂}(a)$.
3. For $N_j = 50.000$ juveniles, a father and mother is drawn with probability proportional to its fecundity.
4. $n_A(a), a = 1, \ldots, 3$ catches are drawn randomly from each of the parent age classes.
5. n_J juveniles are drawn randomly.
6. Find the number of parent offspring pairs in the sample and calculate $\hat{N}_A$ and $\hat{N}_{A_C}$. In the cases where $H = 0$, only $\hat{N}_{A_C}$ is calculated.
7. Repeat this for 50.000 iterations.

The N_j juveniles should be regarded as a sample from a larger population. Assuming births are independent in the process, the number of offspring for each adult is a Poisson variable, and the number of offspring for each adult in a limited sample is multinomial conditional on N_j.

Input data

Input values for the simulation were selected from the RCU population above, and the WWD population from the same data set. The RCU population has a low number of parent-offspring pairs ($H = 7$), and the WWD population was selected as an example of a population with higher number of parent-offspring pairs ($H = 35$). The simulated CKMR's for the RCU population were run with both the variable recruitment model and the more skewed stable population model.

Simulation results

The results of the simulations are summarised in table 3.1–3.4. The two versions of the RCU population model performs very similarly. This is expected, they are both essentially the same Poisson model with $EH \approx 7$. The observed distribution of H in the simulations has a variance close to, but slightly lower than the mean. This is as expected, ref. section 3.1.

If we compare the bias corrected estimator $\hat{N}_{A_C}$ to $\hat{N}_A$, the corrected estimator is very close to unbiased, and have a negative bias of only $< 1.1\%$ for the three simulated populations. The uncorrected estimator has a very large positive bias of 18.9% and 20.5% for the populations with $H = 7$, and 2.6% for the population with $H = 35$. $\hat{N}_{A_C}$ also has as much lower mean squared error and standard error than $\hat{N}_A$. Although the corrected estimator is almost unbiased, this is the result of compensating for the heavy right tail by most of the time underestimating N. The uncorrected estimator has a median value close to the true population size, while the corrected estimator has a median value 10-15% below the true population size.

The CV estimates using the delta method variances near the simulated values for the WWD (H=35) populations, but for the RCU (H=7) population the delta method variances underestimate the CV with 30-35%.

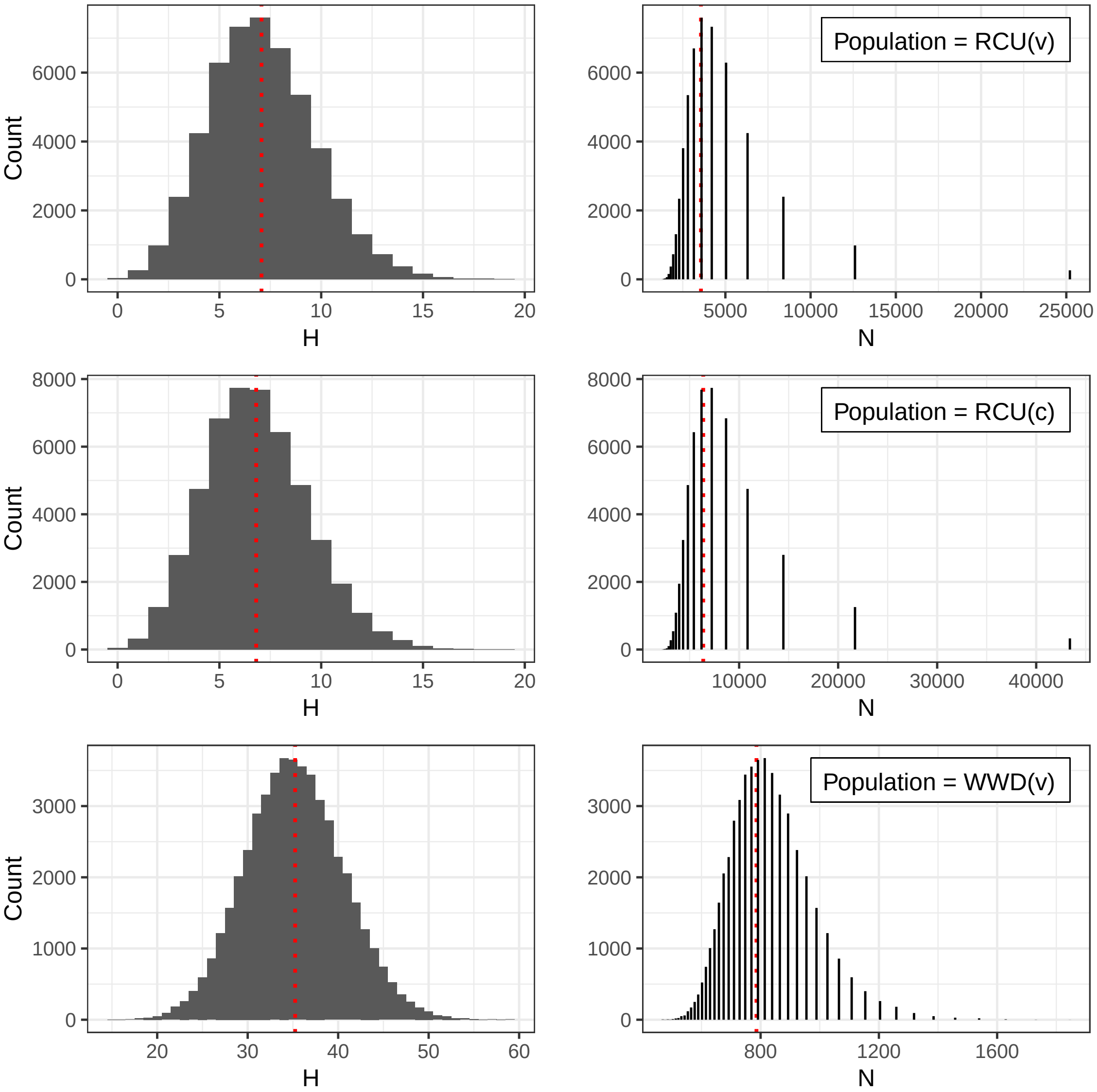


Figure 3.1: Distribution of H (number of parent offspring pairs) and $\hat{N}_A$ for the 50.000 simulations using the population parameters in table 3.1. The dotted red line indicates expected values from model.

.

Table 3.1: Input values to simulations.

		Age distribution			Fecundity			Catches			
	N_A	p_1	p_2	p_3	F_1	F_2	F_3	n_1	n_2	n_3	n_J
RCU(v)	3568	0.287	0.613	0.100	1.00	4.49	7.73	89	78	10	93
RCU(c)	6376	0.682	0.286	0.032	1.00	4.49	7.73	89	78	10	93
WWD(v)	786	0.588	0.191	0.221	1.00	4.49	7.73	100	29	23	107

Table 3.2: Number of parent offspring pairs in 50.000 simulations

		Simulated			Expected	
	N_A	$\bar{H}$	$VAR(H)$	$\overline{1/\sqrt{H}}$	EH	$1/\sqrt{EH}$
RCU(v)	3568	*7.10*	*6.82*	*0.399*	7.07	0.376
RCU(c)	6376	*6.80*	*6.51*	*0.409*	6.81	0.383
WWD(v)	786	*35.21*	*29.59*	*0.170*	35.25	0.168

Table 3.3: Estimated N_A

	N_A	$\bar{\hat{N}}_A$	MED	SD	$\bar{CV}_\Delta$	CV	$RMSE$
RCU(v)	3568	*4243.9*	*3601.7*	*2464.05*	*0.399*	*0.581*	*2555.06*
RCU(c)	6376	*7682.7*	*6200.4*	*4568.32*	*0.409*	*0.595*	*4751.49*
WWD(v)	786	*806.7*	*791.5*	*132.10*	*0.170*	*0.164*	*133.72*

Table 3.4: Estimated N_{A_C} (bias corrected estimator)

	N_A	$\bar{\hat{N}}_{A_C}$	MED	SD	$\bar{CV}_\Delta$	CV	$RMSE$
RCU(v)	3568	*3529.0*	*3151.5*	*1551.64*	*0.288*	*0.440*	*1552.11*
RCU(c)	6376	*6337.2*	*5425.4*	*2887.13*	*0.290*	*0.456*	*2887.37*
WWD(v)	786	*783.3*	*769.5*	*124.21*	*0.161*	*0.159*	*124.24*

Summary of results from the numerical simulations. Output values from the simulations in italics. MED denotes the median value, $\bar{CV}_\Delta$ denotes the estimated $CV(\hat{N})$ from Poisson distribution and delta method.

3.6 Discussion

Bravington et al. (2016b) raises the concern that the pairwise pseudo likelihood approach has limitations for small populations. The focus here has been on developing a CKMR model for an age structured population as an extension to the traditional Lincoln-Petersen estimator for closed populations, and validate its performance, particularly for small populations.

In ordinary mark capture, there is a temporal ordering of the samples and a clear first and second sample. In CKMR, all births in the population is a “mark”, and both the juvenile and adult sample are samples from the same pool of marks. Therefore, it is arbitrary if we take the view that the juveniles mark the parents, or the opposite. Bravington et al. (2016b) takes the viewpoint that the juveniles marks the parent, while Ruzzante et al. (2019) argues that the opposite direction gives a closer relationship to the ordinary Lincoln-Petersen model.

For the CKMR model in this chapter, we model the marking process by weighting all the adults by their expected number offspring (2.19). The simple unweighted CKMR Lincoln-Petersen estimator is easily derived from the offspring marks parent perspective, and the weighted estimator presented here is initially developed from the offspring marks parent point of view and the expected relative reproductive output model. However, the reverse view gives a simpler derivation of the estimator, and gives a more direct interpretation in terms of the Lincoln-Petersen estimator, as discussed in Ruzzante et al. (2019) and chapter 3.1 above. This agrees with the observation by Skaug (2017) that a parent-centric and offspring-centric approach may give different paths to the same results.

By viewing the juvenile population as the marked part of the population, we can think of the sampling process as "hypergeometric with added variance", or more precisely, hypergeometric conditional on the random number of marks. The main benefit of this approach compared to the pairwise pseudo likelihood approach is that we do not need to base the inference on large sample theory, or require sampling to be sparse, but can use the weaker assumptions that the number of pairs is approximately Poisson.

I have here developed a moment estimator closely related to the Chapman estimator, which is nearly unbiased even when the number of recaptures is low. The small population bias of the estimators follow from the inverse relationship between the number of recaptures and population size. It is likely to be a problem with maximum likelihood approaches as well, as it is the case with the ordinary mark recapture model where the Lincoln-Petersen estimator is the maximum likelihood estimator for the hypergeometric model.

The focus here is on the moment estimator, but the principle can be extended to likelihood methods. By viewing the number of marks as a random quantity, it is possible to construct a conditional likelihood given the observed parent's expected reproductive output. By using a hypergeometric or binomial sampling model from the pool of marks, it should be possible to construct a maximum likelihood model for CKMR, even when population size is small and sampling is non-sparse.

3.6.1 Model Assumptions

I will here give an overview of the assumptions of the CKMR and regular Lincoln-Petersen estimator, and how they relate. The main assumptions of the Lincoln-Petersen model, can be summarised as

1. Reliable marking process
2. Closed population
3. Independent samples

Reliable marking process

In ordinary mark-recapture, this is usually formulated as a pair of assumptions, 1. The animal do not lose their tags and, 2. All marks are reported in the second sample. In CKMR accuracy of the identification of parent-offspring pairs is even more crucial. Due to the large number of pairwise comparisons, even small errors in the genetic analysis can lead to a large number of false positives.

In my analysis, I have assumed that the number of parent offspring pairs are reliably determined in the genetic analysis. If it is possible to obtain an unbiased estimate for EH, uncertainty in H could possibly be incorporated into the moment estimator by using the estimated H. However, this would add variance to the model, and inference would have to rely on large sample approximations rather than the Poisson approximation. If the model needs to quantify the uncertainty in parent-offspring pair detection, a pseudo likelihood approach is probably more appropriate.

Closed population

In its strictest form, the closed population assumption says that the population is not subject to recruitment, mortality or migration between the two samples. This is not true for a CKMR model where the population is sampled over a long period, and fortunately the Lincoln-Petersen model allows for several deviations from this assumption (Seber 1982, chap. 3.2.1). The population size might change during the sampling period, but our objective here is to estimate the population size at the time of marking, i.e. when the juvenile cohort is born. We are therefore only concerned with how the deviations from the closed population assumption affects the estimate of N here, not changes in population size during the sampling period.

Juvenile Mortality. In ordinary two sample mark recapture, the Lincoln-Petersen estimator allows for random mortality between the first and second sample, if the mortality affects the marked and unmarked population equally (Seber 1982, chap. 3.2.1). As discussed in chapter 3.1, if mortality is random, the surviving population is a subsample of the original marked population, and a sample from this subsample is identical to a sample from the original population. In our CKMR version of the Lincoln-Petersen model, this applies to the juvenile population, which we regard as the marked population.

Adult mortality. If fecundities can be estimated reliably, it is not required that the adult samples are random samples from the populations. The mortalities do therefore not need to be random with respect to age or size, which, is a major point of the age structured model developed in this chapter. Adults captured before the birth of the juvenile cohort may not live to reproduce, and therefore may belong to the population to be estimated, and must be discounted by the probability that they live to be included into the population.

Recruitment. While emigration is indistinguishable from mortality, and should be treated the same way, immigration poses a bigger problem to the model. If there is migration into the population, juveniles or adults, the mark pool will be diluted, and the population estimates will be too high. If there is a non-negligible amount of immigration, the closed population model fails, and this will have to be taken into account in the population modelling.

Independent samples

A necessary assumption for the Lincoln-Petersen method to produce unbiased estimates, is that the two samples are independent. If some individuals are more likely to be caught than others in both samples, this will cause a negative bias in the population estimates. Similarly, if there is a negative dependence between the samples, this will cause an overestimation of the population size.

There are two causes for correlation bias in mark recapture, 1. local dependence, and 2. individual heterogeneity (Chao et al. 2008). Local dependence arise when the recapture of an animal is affected by the initial capture of an animal, e.g. an animal has a behavioural response and become trap shy or trap happy. Even if the animal is unaffected by the marking process, individual differences in catchability will lead to correlation bias. Sufficient conditions for the Lincoln-Petersen estimator to avoid correlation bias is then either equal catchability for the recapture sample, or if there is no local dependence, equal catchability in either sample (Chao et al. 2008).

This highlights one of the strengths of CKMR compared to ordinary mark recapture, since each animal is only captured once, local dependence is usually not an issue. A counterexample might be if the parents are captured before the birth and, the trapping affects the breeding or survival of the captured individual.

In the CKMR parent offspring model, there are two main sources of heterogeneity (Bravington et al. 2016b). One is heterogeneity in individual capture probability, caused by correlated survival or capture probability between parent and offspring. If there is correlated survival or capture probability, this will increase the probability of recaptures, and deflate the population estimates. While this might arise in CKMR models, for example from weak spatial structures in the population, this again is one the strengths of CKMR. In ordinary mark recapture, avoiding heterogeneity in capture probability is a major issue, since capture probability is usually a function of age, size and other individual parameters.

Reproductive variability The other source, particular to CKMR, is the heterogeneity caused by the variability in reproductive output. If the probability of sampling an adult is correlated with its reproductive output, this will produce a correlation bias, unless accounted for. The main point of the age structured model discussed in this chapter is to account for this heterogeneity in reproduction caused by the age specific fecundities and survival. Even if age and size specific fecundities are accounted for, there can still be individual differences affecting both survival of sampling and fecundities. On an individual level, fecundity and the probability of survival might have a positive or negative correlation, depending on the biology of the species. For example, high status individuals might have both better survival and higher reproductive success, and the true expected fecundity of an individual might be higher than what is predicted by an age or size structured model. This would again increase the number of recaptures and produce a negative bias. On the other hand, if survival is dependent on reproductive effort, the surviving individuals might have a lower expected reproductive output than predicted. As an example, consider an iteroparous salmon species which does not breed every year, and the probability of surviving a breeding season is low. An individual captured one year later would then be likely to belong to the subpopulation that did not breed that year.

For the Lincoln-Petersen model to hold, we then need only to require that the offspring are sampled randomly from the juvenile cohort. There is no requirement on the sampling of adults. Seen from the unweighted offspring marks mother perspective, the offspring sample is not a random sample of the parents. For the Lincoln-Petersen model to hold, it is therefore required that the parent sample is a random sample. This is the approach taken by Rawding et al. (2014) in their Chinook Salmon study, where they use an unweighted CKMR Lincoln-Petersen estimator.

Siblings While it is far easier to obtain a random sample of the equal aged juvenile population than the heterogeneous parent population, it still requires random mixing or careful sampling schemes. Bravington et al. (2016b) notes that a sample should not contain too many littermates, i.e. siblings from the same cohort, and that this might occur even in large sparsely sampled population for example if the juveniles show schooling behaviour and are likely to be caught together.

They also discuss pairwise dependence caused by two parent offspring pairs sharing the same parent. If the juveniles are sampled independently, this type of dependence between siblings are not an issue in our model, since we only require independence of the juveniles conditionally on their parents' fecundity, not pairwise independence.

3.6.2 Demographic model assumptions

The expected relative reproductive output approach to the CKMR model allows for great flexibility of sampling methods and strategies, but requires that we are able to model both the reproductive output of the sampled potential parents, and the whole population.

A basic assumption here is that the population is adequately described by an age structured model. If there are large individual differences in growth within the age classes, a size structured model might give a better description of the population.

Age structure As is seen in the example in section 2.7, the modelled age distribution may have great consequences for the estimated population size. Since the age structure mainly matters for estimating the average fecundity in the population, the importance of accurately determining population age composition increases if there are big differences in reproductive output between the age classes. In the brook trout example, the estimate assuming stable age distribution implies that there is a large undersampled population of the least fecund age class, and therefore gives a much higher relative reproductive weight to the sampled adults than the estimates assuming that sampling is representative.

Age class identification Age of the sampled individuals is important for two main reasons. For the sampled adults, we use age to estimate the expected reproductive output of the sampled individuals. In the brook trout study, age is determined from length, and reproductive output is estimated from a known relationship between size and fecundity. Some misidentifications in the ages of the adults are therefore not likely to affect the estimate much. A large specimen misidentified as belonging to an older age class are also likely to have had a higher reproductive output in the birth year of the cohort.

For the sampled juveniles, age is more critical, since it is needed to determine that they belong to the juvenile cohort born the year in question. While a falsely excluded juvenile will only decrease the sample size, falsely included juveniles may affect the population estimates.

Reproductive output Modelling systematic variations in fecundity is arguably both the hardest, and one the most important aspect of an expected reproductive output CKMR model. Failure to model this accurately may produce biased results. The relationship between age or size and reproductive output can be very complex. It is common to estimate the size specific fecundity from body size, and known relationships between body size and gonad size. The realised relationship between reproductive success and size might differ considerably from this, since size also affect other aspects of the reproductive biology, e.g. reproductive timing, breeding ground competition and ability to defend nests. Even if female reproductive output is well described by the size at age relationship, the males often have higher variability and a more complex relationship. The relationship between male size and reproductive success is usually not dependent on gonad sizes, but rather on competition and reproductive strategies.

A good understanding of the reproductive biology of the species in question is essential in the construction of an expected reproductive output CKMR model. The expected reproductive output approach gives flexibility of sampling, but as discussed in section 3.4, the robustness of the model to errors or uncertainties in reproductive output is highly dependent on the sampling. To reduce bias caused by incorrect assumptions about reproductive output, sampling should be kept as random as possible with respect to age and size.

The CKMR analysis can also provide information about demographic parameters directly. For example, Bravington et al. (2016a) found that the relationship between size and number of offspring was larger than previously thought in southern bluefin tuna by comparing the size distribution of identified parents with the expected number of offspring from the size distribution in the population.

Chapter 4

Half sibling CKMR

Both parent–offspring and half sibling CKMR are only directly informative about the parents. A fundamental difference between the two situations it is that with half siblings, the parents are only observed indirectly through their offspring. In a parent–offspring CKMR, the individual heterogeneity caused by different ages, size and other observable life history characters can be accounted for by using the expected reproductive output framework. In a half sibling CKMR, all individual heterogeneity in the parents is unobserved, and will have to be treated as a latent model. This chapter examine how the demographic heterogeneity affects the probability of sampling a half siblings.

4.1 Same cohort siblings

As noted by Bravington et al. (2016b), a problem with same year siblings that the probability of finding a living sibling is dependent on the variance of the reproductive output. While this complicates the analysis, it does not necessarily make the same cohort siblings unusable for analysis. We should also bear in mind that the variance in reproductive output is an interesting demographic parameter which provides insight into the population processes. In this section I will discuss how the variance affects the probabilities of living siblings in the case where the siblings are born the same year.

4.2 Half sibling probabilities

In the same year case, the siblings are selected from the same reproductive output, and are therefore not independent. Taking into account that an individual cannot be its own sibling,

$$
\begin{aligned}
P(i \text{ and } j \text{ are MHS}) &= \sum_{k \in \mathcal{F}_i} P(i\text{'s mother is } k,\, j\text{'s mother is } k) && (4.1)\\
&= \sum_{k \in \mathcal{F}_i} P(i\text{'s mother is k}) P(j\text{'s mother is } k \mid i\text{'s mother is } k) && (4.2)\\
&= \sum_{k \in \mathcal{F}_i} \frac{F^{(k)}(y)(F^{(k)}(y) - 1)}{B(y)^2}, && (4.3)
\end{aligned}
$$

where $F^{(k)}(y)$ here is the fecundity of female k in the year y, when i and j is born.

Taking the expectation of the above with respect to the fecundity, and assuming approximate independence between the nominator and denominator,

$$
\begin{aligned}
P(i \text{ and } j \text{ are MHS}) &= \frac{N(y)}{B_t^2} E[F(F-1)] & (4.4)\\
&= \frac{N(y)}{B_y^2}\left(E(F^2) - E(F)\right) & (4.5)\\
&= \frac{N(y)}{B_y^2}\left(VAR(F) + (EF)^2 - E(F)\right). & (4.6)
\end{aligned}
$$

Writing EF as $\bar{F}$ and expressing the total number of births as $B = N\bar{F}$,

$$
\begin{aligned}
P(i \text{ and } j \text{ are MHS}) &= \frac{N_y}{\bar{F}^2 N_y^2}\left(VAR(F) + \bar{F}^2 - \bar{F}\right) & (4.7)\\
&= \frac{1}{N_y \bar{F}}\left(\frac{VAR(F)}{\bar{F}} + \bar{F} - 1\right) & (4.8)\\
&= \frac{1}{N_y}\left(CV(F)^2 + 1 - \frac{1}{\bar{F}}\right). & (4.9)
\end{aligned}
$$

If the number of offspring follows a Poisson distribution, $VAR(F)/\bar{F} - 1 = 0$, and $P(i \text{ and } j \text{ are MHS}) = 1/N_y$. The form given in (4.8) gives the probability of being sisters as the average number of sisters for a female offspring, $VAR(F)/\bar{F} + \bar{F} - 1$ (Keyfitz and Caswell 2005, eq. 15.3.3), over the total number of births. This gives an intuitive interpretation of (4.9) as $P(i \text{ and } j \text{ are MHS}) = (\text{average number of siblings/average number of offspring})/N_y$.

In the case where $VAR(F) = E(F)$, $P(i \text{ and } j \text{ are MHS}) = 1/N_y$. If fecundity is high, $1/F$ is small, and

$$P(i \text{ and } j \text{ are MHS}) \approx \frac{1}{N_y}\left(CV(F)^2 + 1\right). \qquad (4.10)$$

It might be easier to apply (4.10) than (4.9), since only the relative fecundities matter, and any property which scales with fecundity can be used.

Juvenile mortality

If the population is subject to random mortality, it will not affect the probability. Let F_0 be the initial number of siblings, and F_t the number of surviving siblings at time t, and s survival probability from 0 to t. The number of surviving offspring is then a binomial variable, with $E(F_t|F_0) = sF_0$ and $VAR(F_t|F_0) = F_0 s(1-s)$. From the law of total variance,

$$VAR(F_t) = s(1-s)EF_0 + s^2 VAR(F_0), \qquad (4.11)$$

so that

$$\frac{VAR(F_t)}{\bar{F}_t} = 1 - s + s\frac{VAR(F_0)}{\bar{F}_0}. \qquad (4.12)$$

Inserting this into 4.8 and cancelling out the s, we see that

$$\frac{1}{N_y\bar{F}_t}\left(\frac{VAR(F_t)}{\bar{F}_t}+\bar{F}_t-1\right)=\frac{1}{N_y\bar{F}_0}\left(\frac{VAR(F_0)}{\bar{F}_0}+\bar{F}_0-1\right). \tag{4.13}$$

4.3 Population structure

All demographic detail wraps into the CV. If most heterogeneity can be assumed to arise from observable traits in the population, for example age and size classes, the CV could be estimated from observed data.

If we can assume that reproduction within the classes are homogenous and random, i.e. reproductive output for individuals is Poisson distributed, the CV can be found by averaging over the groups. Let K be a random variable that denote the group level. A typical example for groups would be age class, but K could equally well refer to size class.

Using the properties of the Poisson distribution, (A.7)–(A.9),

$$F|K \sim Poisson(F_k) \tag{4.14}$$

$$VAR(F) = VAR_K(E(F|K)) + E_K(VAR(F|K)) \tag{4.15}$$

$$= VAR_K(F_k) + E_K(F_k) \tag{4.16}$$

$$E(F) = E_K(E(F|K)) = E_K(F_k) \tag{4.17}$$

$$CV(F)^2 = \frac{VAR_K(F_K) + E_K(F_k)}{E(F_k)^2} = CV(F_k)^2 + \frac{1}{E(F_k)}. \tag{4.18}$$

Setting the CV above into (4.9) gives an expression for the MHS probability as a function of the group specific fecundities and the distribution of groups (e.g. age distribution),

$$P(i \text{ and } j \text{ are MHS}) = \frac{1}{N}(CV(F_k)^2+1) \tag{4.19}$$

$$= \frac{1}{N}\frac{E(F_k^2)}{E(F_k)^2} \tag{4.20}$$

$$= \frac{1}{N\bar{F}^2}E(F_k)^2. \tag{4.21}$$

4.4 Example: A moment estimator

Assume that for a population the age structure and age specific fecundities of the breeding population are determined by sampling. Let F_i be the fecundities of age class i, and p_i be the proportion of the age class.

If we can assume that reproduction is not overdispersed and follows a near Poisson distribution, and that this applies to both sexes. We can apply (4.20),

$$P(i \text{ and } j \text{ are HS}) = \frac{2}{N}\frac{\sum_k p_k F_k^2}{(\sum_k p_k F_k)^2}. \tag{4.22}$$

Let n_J denote the number of juveniles, or potential siblings caught, and H the number of distinct half sibling pairs. If a catch of n_j juveniles, there are $n_J(n_J-1)/2$ ordered pairs, and a moment estimator for N is

$$\hat{N} = \frac{n_J(n_J-1)}{H} \frac{\sum_k p_k F_k^2}{\left(\sum_k p_k F_k\right)^2}. \tag{4.23}$$

4.5 Breeding success

Formulating the half sibling probabilities only in terms of the variance obscures one important issue. Only successfully breeding parents are visible to the analysis, since they are only observed indirectly through their offspring. (4.3) is the same whether all potential mothers are included or not. The following equations, (4.4)–(4.9), also hold whether only successful breeders or all potential mothers are included, although the half siblings clearly only hold information about the successful breeders. To avoid the ambiguity of which females are included in $VAR(F)$ and $E(F)$, it is useful to separate between the successful breeders (breeding population, N_b) and the full population (N). Let $\tilde{F} = F|F > 0$ denote the fecundity of the successful breeders, and $\psi = P(F > 0)$ the probability of breeding success. Using that $EF^2 - EF = P(F > 0)(E(F^2) - E(F)|F > 0) = (VAR(\tilde{F}) + E\tilde{F}^2 - E\tilde{F})\psi$ and setting $B^2 = (\psi N E(\tilde{F}))^2$, and using (4.7)–(4.9) gives

$$P(i \text{ and } j \text{ are MHS}) = \frac{1}{\psi N}\left(1 + CV(\tilde{F})^2 - \frac{1}{E(\tilde{F})}\right), \tag{4.24}$$

as expected. Here $\psi N(y) = N_b(y)$, the breeding part of the population.

4.6 Siblings from different cohorts

The sampling of siblings from different years are a close kin equivalent of multiple sightings of the same animal. Let $S_k(y_i, y_j)$ be a random variable indicating the survival of individual k from time y_i to y_k. We can write (1.15) as

$$P(i \text{ and } j \text{ are MHS}) = \sum_{k \in \mathcal{F}_i} \frac{F^{(k)}(y_i)}{B(y_i)} \frac{S^{(k)}(y_i, y_j) F^{(k)}(y_j)}{B(y_j)}. \tag{4.25}$$

For an illustration of the key difference between within-cohort and between-cohort, I will consider a modification of the constant fecundity and survival model in Bravington et al. (2016b, 3.2). Using the same notation as in 4.5 above, let $\Psi(y_i)$ be a random indicator variable indicating successful breeding in y_j with $\psi = E\Psi$, and $\tilde{F} = F|F > 0$,

$$P(i \text{ and } j \text{ are MHS}) = \sum_{k \in \mathcal{F}_i} \frac{\Psi(y_i)\tilde{F}^{(k)}(y_i)}{B(y_i)} \frac{S^{(k)}(y_i, y_j)\Psi(y_j)\tilde{F}^{(k)}(y_j)}{B(y_j)}. \tag{4.26}$$

If we can assume that survival and fecundity is independent of age and time, but allow and individuals probability of breeding being dependent between years, then taking the expectation and summing over the mothers,

$$P(i \text{ and } j \text{ are MHS}) = N(y_i)\frac{E(\Psi(y_i)\Psi(y_j))E(\tilde{F}(y_i)E(\tilde{F}(y_j))E(S(y_i,y_j))}{N(y_i)N(y_j)E(\tilde{F}(y_i))E(\tilde{F}(y_j))E(\Psi(y_i))E(\Psi(y_j))} \tag{4.27}$$

$$= \frac{s(y_i,y_j)}{N(y_j)}\left(1 + CCV(\Psi(y_i),\Psi(y_j))\right), \tag{4.28}$$

where $s(y_i, y_j) = E(S(y_i, y_j))$ and CCV is the coefficient of covariation (1.4). I.e. correlation between breeding success in y_i and y_j will cause a bias in $P(i$ and j are MHS), and if the breeding events are independent the correlation bias disappears.

In the within-cohort case (4.24), the half siblings contain no direct information about what fraction of the population are breeding, and inference about N would have to make assumptions about $VAR(F)$, the variance in fecundity which also includes the population that did not successfully breed. If we can assume that we have half sibling from two independent breeding events, ψ cancels out from (4.26), and the half sibling samples are informative about the whole population, even if the total fraction of the population which successfully breeds is small.

4.7 Discussion

After this chapter was initially written we discovered a preprint by Akita (2019) (Akita 2018, also.), which arrives at some of the same results as is derived here in (4.2)–(4.4). There are some differences in the formulas, as his model is parametrised in terms of the Negative Binomial dispersion parameter, with his $\phi^{-1} = VAR(F)/E(F)^2 - 1/E(F)$ (see appendix A). The derivation in 4.2 is also found in Wang (2009, eqn 3). However, I also make the approximation $B(B-1) \approx B^2$, which does not appear in the formulation by Wang and Akita.

The main motivation for this chapter was to investigate to what degree half siblings could be used for CKMR analysis, or if the dependence caused by variation in reproductive success makes this impossible. In our case, the intended species for further analysis have been Salmonids, which compared to typical marine species have few and large eggs, and breeding is often limited by breeding ground competition. Emphasis has therefore been more on the variation caused by the difference between the census breeding population and the successful breeding population, rather than by the magnitude of the variation.

Heterogeneity

As with other mark–recapture methods, dependence between the samples caused by variation in capture probabilities may cause a correlation bias. As the parent is unobserved, it is not possible to use parent covariates to explain the variation between mothers directly. To use half sibling pairs in CKMR, we either have to assume independence, use a demographic model, or by some means estimate the variance in fecundity in the populations. I will here discuss three different variants of the independence assumption.

1 All births are independent. If we assume the parent population is homogenous and births are random in the population, the births follow a Poisson process, and the variance is equal to the mean. The variance cancels out from the MHS probabilities, and even if the variance in reproductive output is high, this would not affect the MHS probabilities. If there is known size or age structure in the population, this assumption can be relaxed to independent within each class, as in 4.3.

2 Independent clutches within cohort. In many species, each individual may produce several clutches in a breeding season. It is common with polygynous breeding systems, where the males can fertilise several clutches. Hillary et al. (2018, Supplementary S2.5) discusses this in the white shark case. A male may fertilise several clutches, within-cohort paternal half siblings can therefore be independent, while full siblings are from the same clutch and should be excluded from the analysis. Whether different clutches within a breeding season can be considered independent depends on the biology of the species. If there is mate competition, the number of fertilised clutches may be heavily dependent on size or status, and therefore not independent.

3 Independence between years. The MHS probabilities between siblings from different cohorts are dependent on how fecundity and survival probability changes through the life of individuals. If there are systematic changes in survival and fecundity of mature individuals, it would be necessary to fit a demographic model to (4.25) to take this into account. However, in some species it is reasonable to assume that fecundity is near constant after maturation, e.g. some sharks and mammals (Hillary et al. 2018; Bravington et al. 2016b).

The sources of heterogeneity in reproduction are several. If juvenile mortality is dependent in litter mates, this would lead to heterogeneity in reproductive output, even if the parent population is homogenous. A population with high fecundity, constant or Poisson distributed, and random but high juvenile mortality would have an (approximately) Poisson distributed reproductive output. However, if the mortality is dependent, the reproductive output will follow an overdispersed distribution, and there will be heterogeneity which is hard to model. Typical examples may be where the whole clutch is subject to the same predation on eggs, or where the juveniles lives in family groups.

Individual differences in reproductive output are a result of several processes. The most straightforward to model are processes which are caused by size differences and population structure, where the differences in fecundities may be observed by sampling the population. Another process leading to heterogeneity is mate competition, especially in males. It is very common in territorial species that only high status males get to reproduce, and a high proportion of the male population leaves no or few offspring.

Variance in fecundity

I have attempted here to qualify the statement that for half siblings, the probability of observing a half sibling pair is dependent not only on the expectation, but also on the variance of the reproductive output. It is dependent on the variance in the sense that it is dependent on the heterogeneity in fecundity, not the magnitude of the variance. Here a *homogenous* population is a population where the births are random in the population, i.e. the births follow a Poisson process, and a *heterogenous* population is a population where the birth rates varies between the individuals. In most natural populations, there will be individual differences in reproduction, and the variance will be larger than the mean.

A motivation for the analysis was the question if this variation in reproduction can be used in a CKMR model. Is it possible to estimate the quantity $\phi^{-1} = VAR(F)/E(F)^2 - 1/E(F)$, and use this in a CKMR analysis? It follows from (4.13) that ϕ is unchanged under random mortality. It is clear that this cannot be estimated from the number of half sibling pairs alone, since there are two unknowns. A possible approach could be to use the number of higher order sibling groups, i.e. the number of sibling triads and higher, which is also dependent on the dispersion parameter, for example by fitting a zero truncated Negative Binomial. The result would, however, be very dependent on the choice of model. As the number of non-successful breeders are unknown, we see from (4.24) that any number of zero inflated models could fit the observed data, since there is no information about ψ in the sampled juveniles. This is a general problem with estimating the ϕ^{-1} from either the successfully breeders, e.g. from kinship analysis of the offspring, or from the individual variation in fecundity as measured by for example egg count. The estimated variability would only be for the successful breeders, i.e. $\tilde{\phi}^{-1} = CV(\tilde{F})^2 - 1/E\tilde{F}$ in (4.24), while a CKMR abundance estimate would need to use the variance which included the part of the census population which do not successfully breed.

The connection between the half sibling probabilities and reproductive variation is closely related to the effective breeding size. Wang (2009) uses the relationship between reproductive variance and the probability of sampling a pair of siblings as basis for his method for estimating effective population size. For a half sibling CKMR analysis, we could ask what are the requirements for half siblings to be informative about the census population size, and not only the effective number of breeders. As discussed above, there are two main options. Either to use only the half siblings from independent breedings, or to estimate the correlation bias caused by the heterogeneity in reproduction. To estimate the correlation bias, we would need to estimate the variance in reproduction, which is also the main variable determining the effective breeding size/census size ratio, N_b/N.

Chapter 5

River Etne Atlantic salmon kinship analysis

The Atlantic salmon is a species of great economic and conservation value in Norway, which have experienced a decline in abundance in the last decades. The total Norwegian salmon population consist of more than 400 subpopulations in rivers along the Norwegian coast (Forseth et al. 2017), and a validation of the use of CKMR for river salmon population could potentially be of utility for the management of salmon populations.

The data set examined is from the River Etne, where an upstream migration trap has been in operation since 2013. The trap samples the majority of returning adults entering the river each year (Skaala et al. 2015). This near complete sampling of adults provides a time series of absolute abundances, and is a potentially suitable system for validating the CKMR method. I will examine the data from Harvey et al. (2017b,a), consisting of the data from salmon trapped in 2013. As the current data set is a single year only, parent-offspring pairs are not expected, and the focus of the analysis is on half siblings.

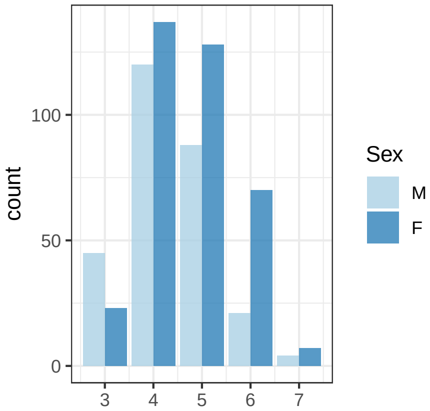
The purpose of this chapter is to give an initial analysis of a data set of Atlantic salmon to determine if it suitable for use as a case study for a half sibling CKMR model.

5.1 Data set

The data are collected in the salmon trap in the River Etne, in Hordaland county of southwestern Norway. The trap is Resistance Board Weir installed near the river mouth. The majority (>90%) of all adult fish are caught entering the river. The caught fish are measured (weight and length), a tissue sample from the adipose fin is taken for later DNA analysis, and scale samples (2-4 scales) are taken. Fish identified as escaped farm salmon are killed, other fish are released to continue upstream migration. The data set analysed here is the caught adults from the 2013 season, previously used in a study by Harvey et al. (2017b), where a further description of the site and genotyping procedures can be found. The data set is available in a Dryad repository (Harvey et al. 2017a). All specimen are genotyped using 31 microsatellite loci. Scale readings have been used to find smolt age, sea age and year of previous breedings

Table 5.1: Number of wild salmon specimen in each sex and age class.

Age	Males	Females	Unknown	Sum
3	45	23	4	72
4	120	137	9	266
5	88	128	7	223
6	21	70	1	92
7	4	7	0	11
NA	218	246	10	474
Sum	496	611	31	1138



for the entire data set. The age readings are not complete, as not all sampled scales can be used to determine the both smolt and sea age. All scale readings contain two binary attributes to indicate the certainty of smolt age and sea age (certain/uncertain).

5.2 Life history and demography

The Atlantic salmon is an anadromous species with highly variable life histories, both within and between populations. Eggs are laid in the gravel in rivers (redds). After 1–5 of years in the river, at the size approximately 11–12cm the juveniles smoltify and migrates to the sea. In the 2013 data set, smolt age of the returning salmon ranged from 2–4, with an average of 2.5 years among the specimen where smolt age was determined. Part of the male population will mature as parr before smoltification, and try to breed before migrating to sea. The salmon will then feed and grow before returning to the river to breed. Some individual will mature and return after one year (grilse), but most will feed and grow for 2–4 years before reaching maturity and returning. After breeding, the adults return to sea, where they will feed and grow for 1–2 years before returning to breed again. Most repeated breeders will have a rest year without breeding, but some will return to the river after one winter. In the 2013 data set, of the 100 repeated breeders where previous breedings where readable from scales, 24 returned after one year and 76 returned after two years.

The breeding system is polygamic for both sexes. Fleming (2011) gives a review of salmon breeding system. A female deposits her eggs in a number of nests, depending on her size, which will be fertilised by one or more males. The male breeding population is a mixture of returning anadromous males, and precocious mature parr. The reproductive success of the anadromous males increases with size, but for the mature parr there is no clear size dependency. For the females, the reproductive success increases more with size than is expected by egg count alone. The total sex ratio of the breeding population is skewed towards males because of the precocious male population. The adult sex ratio varies both between rivers and through the breeding season.

5.3 Kinship analysis

From the demography of Atlantic salmon, we should expect the data set to contain a number of same cohort half and full sibling pairs. Parent offspring pairs are possible, but not likely because of the long maturation period. As there is usually a rest year between multiple breedings, there should be a higher proportion of half siblings separated by two years than one year. Half siblings are indistinguishable from grandparent-grandchild and aunt-nice pairs using standard kinship analysis, but the generation time should preclude grandparent-grandchild pairs and make aunt-niece (or male equivalent) pairs very rare compared to half siblings.

In a pairwise analysis of kinships, the expected number of related pairs is only a very small fraction of the total number of comparisons. It is therefore necessary to have control of both the number of false positives and negatives to be able to estimate the true number of related pairs in a sample. Define the LOD score as the likelihood ratio of half sibling over unrelated (Skaug et al. 2010),

$$LOD(i,j) = log\left(\frac{P(g^{(i)}, g^{(j)}|\text{half siblings})}{P(g^{(i)}, g^{(j)}|\text{unrelated})}\right). \tag{5.1}$$

Where $P(g^{(i)}, g^{(j)}|\text{half siblings})$ and $P(g^{(i)}, g^{(j)}|\text{unrelated})$ is the probability of observing a genotype pair $(g^{(i)}, g^{(j)})$ given that they are half siblings or unrelated respectively. The probabilities of a genotype pair under the different kinship hypothesis is calculated assuming Hardy-Weinberg equilibrium and linkage equilibrium (Bravington et al. 2016b, eq. 5.1 and 5.2).

Bravington et al. (2016b) section 5.1–5.2 give a two-step procedure for handling errors in genotyping. The first step is to find a LOD cutoff point high enough that false positives can be ignored. The second step is to use the definite pairs to estimate the false negative rate, using the expected distribution of the half sibling/unrelated LOD distribution. If the data set also contains full sibling pairs, they need to be accounted for to be able to estimate the LOD distribution. Hillary et al. (2018, supp. S2) and Bravington et al. (2017) discuss procedures to account for full siblings and estimate the half sibling/unrelated LOD distribution.

The data set contains 49 individuals with missing data at some loci. As this is only intended as an initial analysis, these were rejected and is not part of the LOD analysis, reducing the sample size to 1089 individuals. I used the *familysim* function in the R Package *Related* (Pew et al. 2015) to generate a bootstrap simulation of pairs of unrelated, half siblings and full siblings under assumptions of Hardy-Weinberg equilibrium and linkage equilibrium. The allele frequencies used for the simulations were calculated from the full data set of 1138 individuals.

The simulated distributions in fig. 5.1 show that the half sibling distribution is very close to the unrelated, and that a cutoff to exclude false positives would have to very close to the expected mean of the distribution. The full sibling distribution also has a large overlap with the half sibling distribution, which would make it very difficult or impossible to remove the full sibling pairs from the analysis to estimate the true distribution of half siblings using only the half sibling/unrelated LOD distribution. If we compare histograms of the expected distributions to the LOD scores from the pairwise comparisons from the salmon data set, the expected modes from the full and half sibling are not identifiable in the tail of the LOD distribution (see fig.

5.1 and 5.4). The tail is also much heavier than expected from unrelated, which is likely to be caused by the presence of a large number of relatives with a weaker relationship than half sibling. Given the demography, the population is expected to contain a large number of first cousins. Half first cousins should be most common, but we should also expect to see full first cousins. Except for the heavy tail, the LOD score for unrelated is very close to the expected. The skew in the distribution seen in fig. 5.3 is very small compared to the total number of pairs in that range, as can be seen from fig. 5.1. Some deviations from the theoretical distributions are expected. Because of genetic linkage, the loci are not completely independent, and the genotype LOD score calculated here by summing over the individual loci is really a pseudo likelihood, not a true likelihood Hillary et al. (2018, supp.).

Using the genetic data in the salmon data set, it is possible to set the LOD score high enough to identify a number of half siblings. However, based on the lack of separation both between the half sibling LOD score for unrelated and true half siblings, and between half and full siblings, it is not feasible to use the LOD distributions to find a usable estimate of the false negative rate.

As a visual tool in the analysis of the sibling matching, I used the software *igraph* (Csardi and Nepusz 2006) and *ggplot* (Wickham 2016) packages for R (R Core Team 2018). When the cutoff is high enough to prevent false positives, the siblings are ordered into clusters, otherwise there will be many random connections between the nodes (fig. 5.6). The connection between age and probability of sibship can also be investigated using these plots. If the sibling clusters are primarily of same age, this provides a validation of both the ageing and sibships. If sibling groups are a mixture of different ages, this indicates that most likely there are errors in the ageing or in in the kinship.

Table 5.2: LOD mean and standard deviation for the simulated pairs. (See fig. 5.1)

	mean	sd
U	-5.73	2.99
HS	7.05	4.10
FS	16.74	4.49

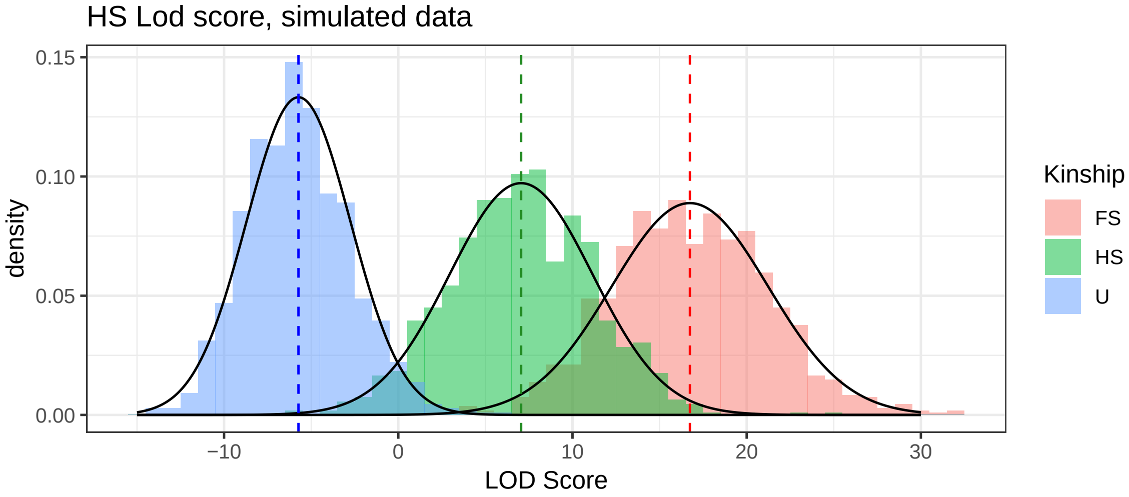


Figure 5.1: LOD score for pairs of full siblings (FS), half siblings (HS) and unrelated (U) for 1089 randomly generated pairs each kinship type. Pairs generated with *familysim* using allele frequencies from the salmon data set. The black curves are Gaussian approximations using the mean and standard deviation from the bootstrapped samples (tab. 5.2). Dotted vertical lines show the mean of each distribution.

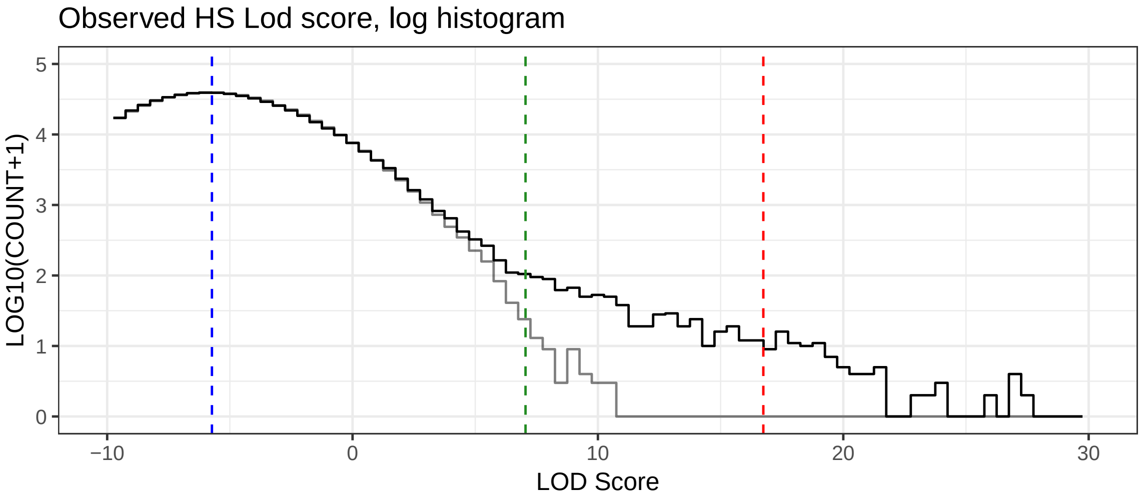


Figure 5.2: Log histogram for LOD scores. The black line is log counts for salmon data. The grey line is simulated unrelated data (all pairwise combinations, 1089 simulated individuals). Total $1089(1089 - 1)/2$ pairwise comparisons for both actual and simulated data. Blue, green and red lines are means from expected distributions (fig. 5.1).

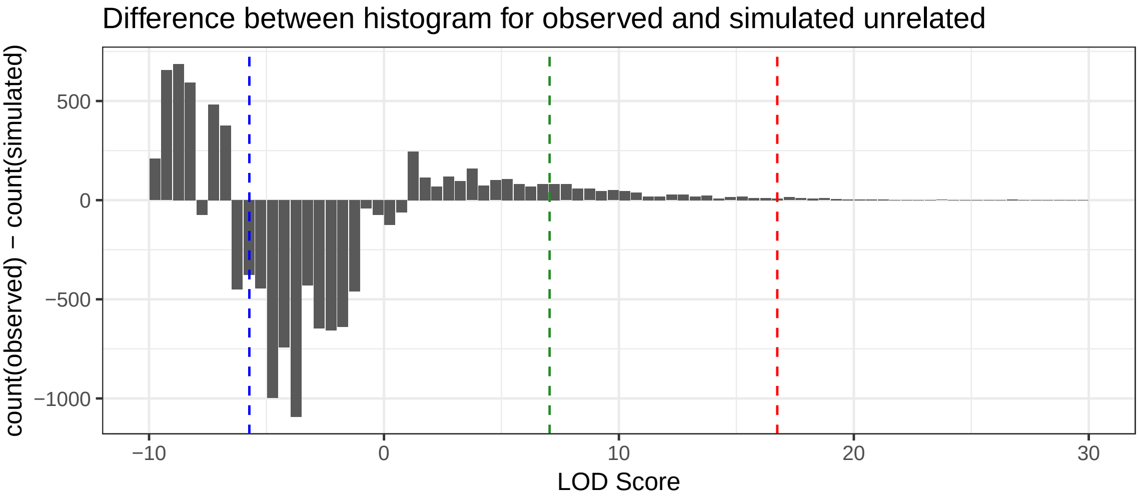


Figure 5.3: Difference between the counts in each bin for observed values and simulated values for unrelated pairs. The observed distribution is skewed left of the expected distribution, and has a heavy right tail. Blue, green and red lines are means from expected distributions (fig. 5.1).

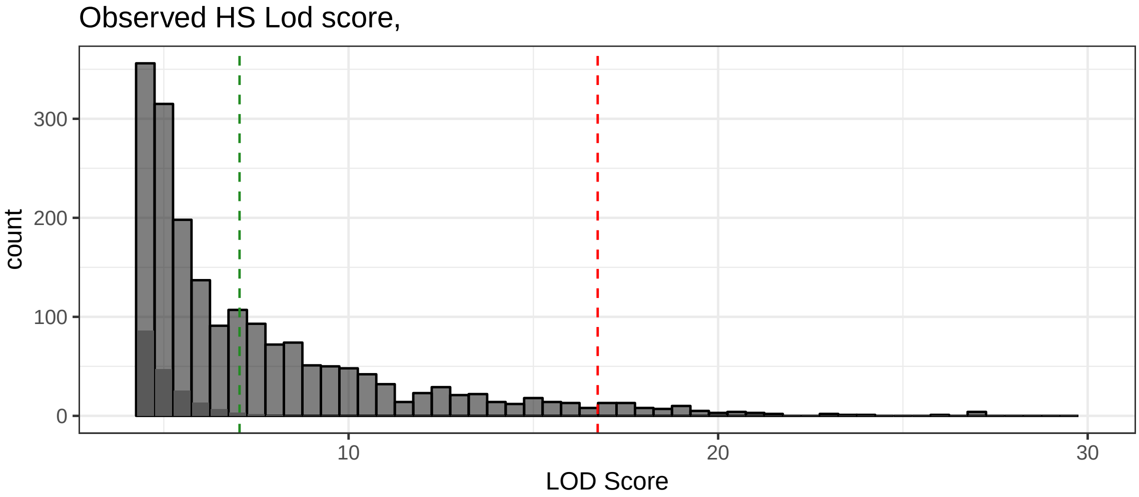


Figure 5.4: Right tail of histogram of observed pairwise LOD scores. The darker columns to the left is expected number of unrelated from Gaussian approximation of unrelated LOD (the solid line in fig. 5.1)

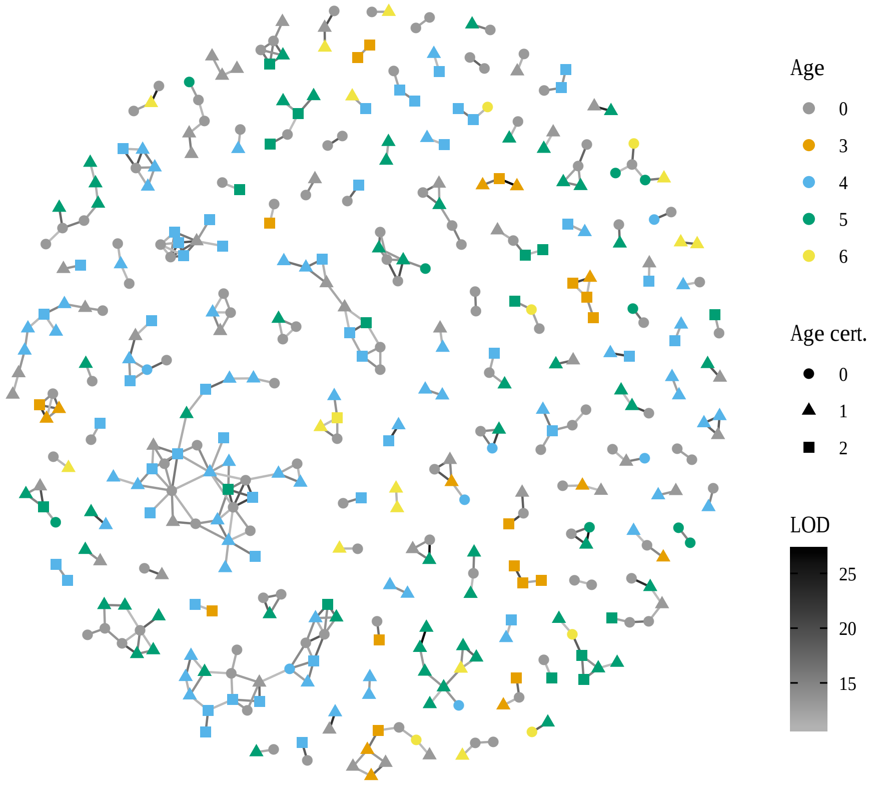


Figure 5.5: Sibling clusters from pairwise kinship analysis with LOD Score > 7. *Age cert.* shows indicated certainty of scale reading for smolt and sea age, 2=both certain, 0=both uncertain. Edges indicate certainty of kinship (LOD).

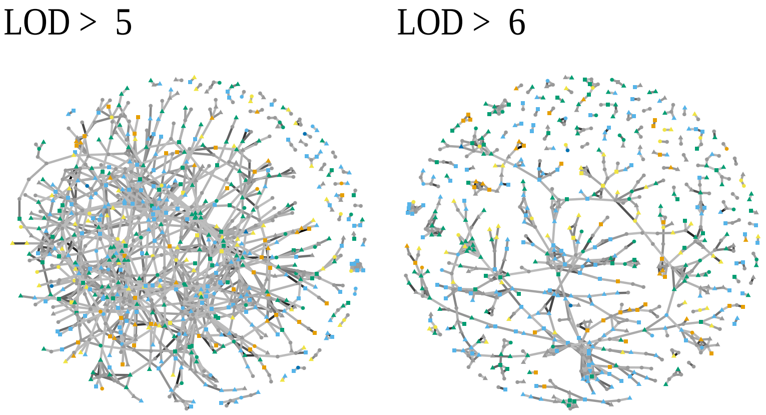


Figure 5.6: Example sibship graphs from pairwise analyses where the LOD Score cutoff is not set high enough to avoid false positives.

Sibling assignments from COLONY

An alternative to pairwise kinship analysis is pedigree reconstruction methods, which use the likelihood of the pedigree structure, rather than pairwise likelihoods. With sibship reconstruction methods, the accuracy of sibship detection increases with the number of individuals in the sibling groups sampled. Sibship reconstruction methods can be highly accurate, but are less successful when cousins and other less related individuals are included in the sample (Städele and Vigilant 2016). For initial exploratory analysis of the sibship structure in the population, I used the software COLONY 2 (Wang 2009; Jones and Wang 2010), which implements pedigree reconstruction for polygamous populations. According to Wang (2009) the COLONY2 algorithm is very accurate for sibling groups, even with few markers, if the sibship sizes are large. This should give a more accurate identification of the siblings which belong to sibling groups. For the siblings belonging to a single unconnected sibling pair, the pedigree analysis has no additional information to infer a parent's genotype prior, and the ability to identify these are not expected to differ substantially from pairwise analysis.

The data set was analysed using COLONY V2.0.6.5. Colony was run with male and female polygamy, long run, high precision and full likelihood. COLONY assigns an uncertainty estimate to each assigned pair in the form of a probability, based on the Monte Carlo simulations used in the pedigree inference. Summing these will give an estimate of the total number of sibling pairs, but the bias and other properties of this estimator is unknown. An issue with using COLONY and similar software for half sibling CKMR, is that the methods are not developed for CKMR in particular. COLONY makes no attempt to estimate the bias or uncertainty in the number of true kinships in the population, which is the uncertainty estimates needed for CKMR analysis. The number of relationships assigned by COLONY is given in table 5.3.

Table 5.3: Number of kinship pairs assigned by COLONY in the Etne River salmon data set. P is the probability of each relationship assigned by COLONY. ΣP is the sum of the probabilities, which give an estimate true number of pairs.

	COUNT($P > 0.9$)	ΣP
Full siblings	118	130
Half siblings	1078	1445
Parents	0	0

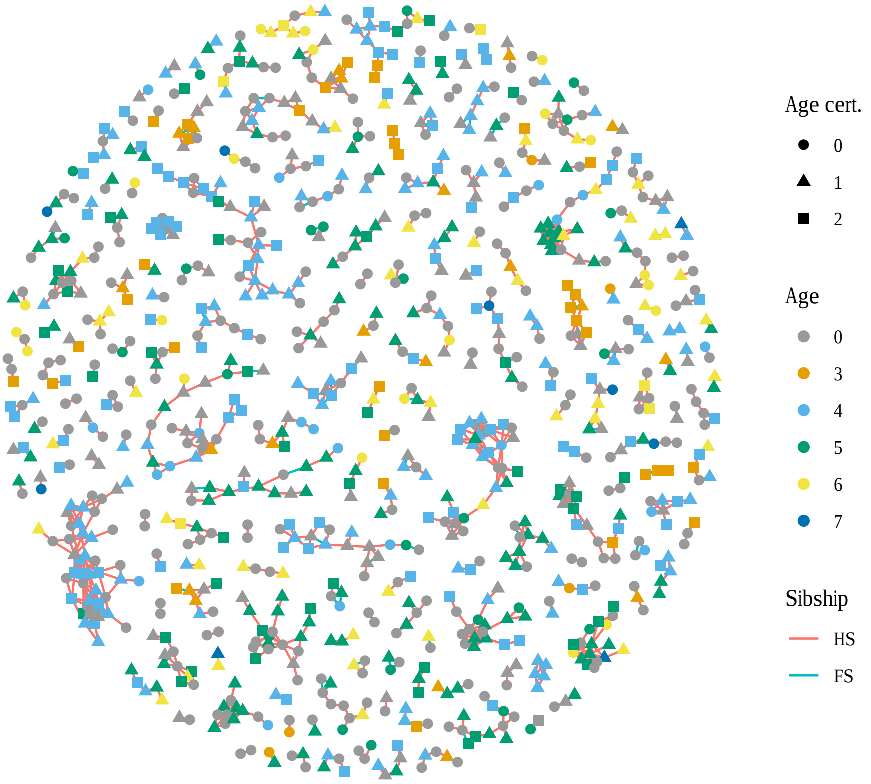


Figure 5.7: All sibling clusters assigned by COLONY with probability > 0.9.

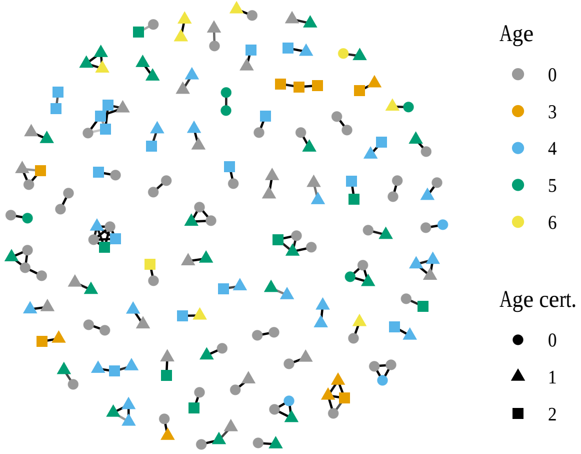


Figure 5.8: Full sibling clusters assigned by COLONY with probability > 0.9.

5.4 Discussion

Is the salmon data set suitable for a half sibling CKMR analysis? The genetic markers used in the data set are, as discussed above, not powerful enough to estimate the true number of half siblings in the sample. A relevant question is still whether the population a suitable model population for validating a CKMR, if more markers were available. There are substantial differences in both the breeding sex ratio and demography of males and females in salmon. A requirement for the construction of a meaningful demographic model for half sibling CKMR is to be able to determine the sex of the parent.

Same cohort siblings

As discussed in chapter 4, using same-year half siblings to estimate the census population size require either using only siblings from independent breeding events, or by some means estimate the variance in reproductive output. Salmon reproduction is highly size dependent, and cannot be regarded as independent, neither between individuals or between half siblings from different redds.

Even if it is possible to make an estimate of the variation in reproductive output among the successful breeders, for example from demographic considerations, or from sibling count in each sibling group, this would likely not be sufficient to estimate the census population size in this case. To estimate census size from half siblings, we would need to estimate the variation in reproductive output which includes not only the successful breeders, but also the part of the census population which did not successfully breed. Because of breeding ground competition, the probability of successful breeding will decrease with increasing population. Even with knowledge of the reproductive variability in the breeding population, we would only be able to estimate the size of the breeding population, not the full census population. This is a consequence of the type of competition. If, for example, competition only affected offspring survival rates rather than breeding probability, (4.13) would apply and ψ could be independent of N.

Different cohort siblings

In a species where fecundity is constant, half siblings from different cohorts can be used to estimate mortality. In salmon, both growth and previous breeding history will affect the probabilities of observing a half sibling. All variables in (4.26) will affect the probabilities. Growth will increase the fecundities, and previous breeding history will affect the breeding probabilities (Ψ), as most adults will have a rest year between the breedings.

Using different cohort siblings would also require a careful consideration of the ageing errors. Since there are many more same-year siblings than different year siblings, it follows from Bayes' theorem that the siblings determined to belong to different cohorts from the scale readings will have an increased error rate.

Conclusion

With the current set of genetic markers, the data set is not suitable for a half sibling CKMR analysis. The markers are not informative enough to identify the half siblings with sufficient power. There is also no information of the parent's sex in the markers, which would be crucial for half sibling based CKMR in a species where there are substantial differences in the male and female demography.

5.5 Possibility of parent-offspring CKMR

A parent-offspring CKMR analysis for this salmon population would be more straightforward, and avoid most of the problems of half sibling analysis discussed. Genetically, parent-offspring pairs are easier to identify with certainty, and genetic analysis can be conducted using standard genetic pedigree software such as COLONY, as is done for the brook trout populations analysed in chapter 2.

In the parent-offspring model discussed chapter 2 and 3, emphasis have been on parents sampled after the birth of the offspring. The pre-breeding mortality for brook trout sampled before the breeding season have been estimated and used as an input to the CKMR estimator (2.21). If we were to perform a parent-offspring CKMR analysis of a salmon population where the potential parents are sampled upon entering the river, they are subject to a mortality from river fisheries before the breed. Since this case differs from the previously discussed case, I will give some preliminary discussion here.

Let $N_{R♀}$ be the size of the female population entering the river, and assume that the population is divided K classes of distinctive size or age. Let F_k,S_k and p_k be the class specific fecundity, pre-breeding mortality and frequency. The expected reproductive output of an individual in class k is then S_kF_k, and the total reproductive output from females of all classes is $\sum_k N_{R♀}p_kS_kF_k$. If j is an offspring born in the year i was sampled, we have from (1.14)

$$P(i \text{ is } j\text{'s mother}|k) = \frac{S_kF_k}{N_{R♀}\sum_l p_lS_lF_l}. \tag{5.2}$$

If the adults entering the rivers are sampled randomly with respect k,

$$P(i \text{ is } j\text{'s mother}) = \frac{1}{N_{R♀}}. \tag{5.3}$$

Which means that, provided we have a random sample of the adults entering the river, the simple Lincoln-Petersen CKMR estimator will give an estimate of the population entering the river, before river fishery mortality. This holds even if the population is subject to size selective mortality before breeding. This is in contrast to the CKMR performed on parents *after* breeding, which is informative about the population size at the time of breeding only, regardless of when the sampling of the parents is done.

A simple Lincoln-Petersen estimator for the total immigrating female population in the year y can then be found as

$$\hat{N}_{R♀}(y) = \frac{n_{R♀}(y)n_J(y)}{H(y)}, \tag{5.4}$$

where $n_{R♂}(y)$ is the number of sampled immigrating parents, and $n_J(y)$ is the number of sampled juveniles born in year y.

The mature parr is still problematic, since we know nothing about the relative contribution of the precocious parr. If we write $R_{parr} = N_{parr}\bar{F}_{parr}$ for the number of offspring fertilised by mature parr,

$$P(i \text{ is } j\text{'s father}|k) = \frac{S_k F_k}{N_{R♂} \sum_l p_l S_l F_l + R_{parr}}. \tag{5.5}$$

To avoid the problem of mature parr, a parent-offspring CKMR analysis can be performed on the female parents only, and the male population size could then be estimated from the sex ratio of the immigrating population.

Age information is less of a problem for a parent-offspring analysis where the parents are sampled while migrating upstream. A matching offspring is likely to have been born the year the sample were taken, and not in the year before or after, since females will rarely breed again without a rest year. We also need the total number of sampled offspring born in the same year, $n_J(y)$ to apply (5.4), but this should be possible to estimate from the distribution of the successfully aged individuals.

The simplicity of the approach above follows from having a random sample of the breeding population. If the parent-offspring analysis is performed on the adults leaving the river after breeding, or later, it would allow for lethal sampling. However, the assumption that the adult sample is a random sample of the reproducing population might then no longer be valid, and a demographic model is needed to describe the expected reproductive output.

Data from The River Etne facility could be a very suitable as a case for a salmon parent-offspring CKMR, when samples from more years are available genotyped and aged.

Chapter 6

Some results in continuous time

In this chapter I use the continuous demographic framework, rather than the discrete framework used in the previous chapters. Continuous demographic models are often used by human demographers (Keyfitz and Caswell 2005; Vaupel and Missov 2014, e.g.), and the discrete models are usually preferred by animal demographers (Caswell 2001, e.g.). The main advantage with the continuous time models is that is often possible to arrive at closed form solutions. Another advantage is that it is often possible to specify models for the whole life history of animals using parametrised models with few parameters.

This chapters builds on the theory developed in Skaug (2017). In particular his decomposition (1.18) is the motivation behind this chapter in a CKMR context. The first part gives a review of stable age theory for continuous demography and the derivation of the equations for probability of living parents, with individual heterogeneity added explicitly to the equations. The second part discusses the case where the age of both offspring and parent are unknown, and gives several new extensions to Skaug's result that if fecundity is constant, the probability that a random individual has a living parent is 1/2. Finally, the third part discuss the case where the survival of parent and offspring is not independent in a stable age distribution context.

6.1 A review of stable age theory

One of the most fundamental results in demographic theory is that of the stable age distribution, first derived by Euler for the discrete case, and later by Lokta for the continuous case. The theory states that a population closed to migration with vital rates unchanging over time, will converge to a stable age distribution, where the population grows at an exponential rate.

There are several ways to derive the stable age distribution. The discussion below follows in Keyfitz (1968, chap 5.1 and 5.2), but take care to see how and if individual heterogeneity affects the results, using basic concepts from frailty theory. Demographic notation mostly follows Keyfitz and Caswell (2005).

Generally, basic demographic parameters will differ between the sexes, and in many cases will be determined by the female part of the population alone. I will follow the common approach of referring to the female part of the populations, unless stated otherwise.

Table 6.1: Notation for continuous time demographic models, adopted from Keyfitz and Caswell (2005).

a, x	age
α	age of maturity
ω	maximum reproductive age
t	time
$N(t)$	total population at time t
$B(t)$	total number of births at time t
$m(x)$	number of offspring at age x
$\bar{b}$	average birth rate in population
$g(x)$	age distribution of females giving birth
$l(x)$	probability of survival to age x
$\mu(x)$	mortality rate
$M(x)$	cumulative hazard, $\int_0^x \mu(a)da$
$f(x)$	age distribution
$\phi(x)$	net maternity function $l(x)m(x)$
r	population growth rate
z	phenotypic unobserved parameter, e.g. frailty
$\pi(z)$	distribution of phenotype Z at birth
$\mathcal{P}$	Probability that a randomly selected individual has a living mother
j, i, h	daughter, mother, grandmother (if used standalone)

6.1.1 Survival

We define survival function $l(a)$ as the probability of survival for an individual to age a. However, in many cases the survival probabilities may vary considerably between individuals. Using terminology from frailty modelling, let Z denote some random variable, discrete or continuous, which is assigned to individuals at births according to some distribution $\pi_0(z)$. This may be an unobserved parameter, or frailty variable, but could also refer to an observable parameter like growth rate. The notable assumption here is, like with the other vital parameters, that the distribution of frailty at birth, $\pi_0(z)$ is constant over time. Using this, we can write $l(a|z)$ for survival of individuals with frailty z, and

$$\bar{l}(a) = E_z(l(a|z)) \tag{6.1}$$

for the cohort survival function. The relationship between individual survival functions and cohort survival are the subject of frailty theory. A compendium of useful relations is found in Vaupel and Missov (2014).

6.1.2 Reproduction

The reproductive rate, or fecundity, is here defined as $m(a)$ if a female with probability $m(a)da$ will produce a single offspring in the time interval da. As with survival, the reproductive function may vary considerably between individuals. The net maternity function, $\phi(a) = l(a)m(a)$ is defined as the expected reproductive output at age a for a newborn individual.

Where the reproductive function is defined through the probability of having an offspring in the age interval $[a, a + da]$ conditional on survival to age a, the net maternity function is the unconditional probability for a newborn of having an offspring in the same interval. Allowing for both fecundity and survival to depend on the random variable Z above, we can find the marginal, or cohort, net maternity function as

$$\bar{\phi}(a) = E_z(l(a|z)m(a|z)). \tag{6.2}$$

The total expected reproductive output of a newborn female over her lifetime is denoted R_0,

$$R_0 = \int_0^\infty \bar{\phi}(a)da. \tag{6.3}$$

6.1.3 Renewal equation

Let $B(t)$ denote the total number of births at time t. The number of births at time t due to the cohort born a years ago is then $B(t-a)\bar{\phi}(a)da$. The process will also have births due to the females present at the start of the present, which gives the initial conditions for the, denoted $G(t)$. Integrating over all reproductive ages, the total number of births at time t is

$$B(t) = G(t) + \int_\alpha^\omega B(t-a)\bar{\phi}(a)da, \tag{6.4}$$

where α and ω is minimum and maximum age of reproduction. Since $\phi(x)$ is zero outside this interval, the integral may be taken from 0 to ∞. The term $G(t)$ is zero when all the initial females are past reproductive age, i.e. $t > \omega$, and the equation is often given with the $G(t)$ term omitted. This equation is known as the renewal equation, or the Lotka integral equation.

6.1.4 Characteristic equation

Considering (6.4) with the initial conditions $G(t)$ omitted, it can be seen by substituting $B(t) = Qe^{rt}$ that this is a solution. Inserting this and simplifying, we get the characteristic equation, or Euler-Lotka equation:

$$\int_0^\infty e^{-ra}\bar{\phi}(a)da = 1. \tag{6.5}$$

Since $\psi(r) = \int_0^\infty e^{-ra}\bar{\phi}(a)da$ is a decreasing function of r, with $\psi(-\infty) = \infty$ and $\psi(\infty) = 0$, this equation can only have one real root. The general solution to (6.4) is $B(t) = \sum Q_i e^{r_i t}$, where the possible non-real roots come in conjugate pairs (Keyfitz 1968). The real root is larger than the real part of all the complex roots, and will dominate the populations dynamics, so that this root will give the long-term growth rate of the population. The Q's may be chosen to satisfy the initial conditions $G(t)$. The complex roots have little impact on the long-term dynamics, but will affect the time to convergence, and introduce waves in the age distribution of the initial generations.

6.1.5 Stable age distribution

The proportion of females aged a, $f(a)$ in a population must be proportional to the surviving part of the cohort born a years ago, so that $p(a) \propto B(t-a)\bar{l}(a)$. Assuming the effect of the initial population has worn off, $f(a) \propto B(t)e^{-ra}\bar{l}(a)$. Normalising this we find the stable age distribution of the population:

$$f(a) = \frac{e^{-ra}\bar{l}(a)}{\int_0^\infty e^{-rx}\bar{l}(x)dx} = e^{-ra}\bar{l}(a)\bar{b}, \tag{6.6}$$

where $\bar{b}$ is the average birth rate in the population. To see that $\bar{b}$ is the birth rate, express $N(t)$ as the composition of all the surviving cohorts, $N(t) = \int_0^\infty B(t-x)\bar{l}(x)dx$. With stable age distribution,

$$N(t) = B(t)\int_0^\infty e^{-rx}\bar{l}(x)dx \tag{6.7}$$

and

$$\bar{b} = B(t)/N(t) = \left(\int_0^\infty e^{-rx}\bar{l}(x)dx\right)^{-1}. \tag{6.8}$$

Age distribution of mothers

The stable age argument can also be applied to the age of mothers at births. The number of births by the cohort aged a is $B(t-a)\bar{\phi}(a) = B(t)e^{-ra}\bar{\phi}(a)$. The normalisation constant here is the characteristic equation, and the age distribution of birthing mothers becomes

$$g(a) = e^{-ra}\bar{\phi}(a) = e^{-ra}E[l(a)m(a)]. \tag{6.9}$$

Stationary populations

In the case of stationary populations, $r = 0$, and the stable age distribution have the simpler form $f(a) = \bar{l}(a)/\int_0^\infty \bar{l}(x)dx$. The normalisation constant here is the life expectancy of a newborn female, since by the tail probability formula we have that $\int_0^\infty \bar{l}(x)dx = \int_0^\infty P(T_d > x)dx = E(T_d)$.

When assuming a stationary population, there is also an assumption of density dependence, stated or unstated. A population may be regulated by both the birth rate and mortality rate. In the cases where I assume a constant population, the life table is kept constant while the fecundity is adjusted, i.e. density dependence is assumed to work on the number of births, not mortality.

6.1.6 Mother-daughter probabilities

Cohort view

The probabilities of living ancestors may either be derived from the cohort perspective, by what Keyfitz and Caswell (2005, chap. 15.1) call the counting method, or from the individual view by what they call the Probability method. The derivation below is based on the counting method.

This assumes independent regimes for survival for mother and daughter, and independence between reproduction and survival. These assumptions will be relaxed later.

The number of living daughters born by the cohort aged x at time t is then $B(t-x-a_j)\bar{l}(x)\bar{m}(x)\,dx\,\bar{l}(a_j)$. A fraction $\bar{l}(x+a_j)/\bar{l}(x)$ of the cohort of mothers will further survive to time t. The total number of living mother-daughter pairs is then

$$N_m(t,a_j) = \int_\alpha^\omega B(t-a_j-x)\bar{l}(x)\bar{m}(x)\frac{\bar{l}(x+a_j)}{\bar{l}(x)}\bar{l}(a_j)dx. \tag{6.10}$$

The number of living daughters belonging to the daughters cohort is

$$N_d(t,a_j) = B(t-a_j)\bar{l}(a_j). \tag{6.11}$$

Using the stable age assumption, $B(t-a_j) = e^{-ra_j}B(t)$ and $B(t-a_j-x) = e^{-r(a_j+x)}B(t)$. Inserting this, we can find the probability of having a living mother as the ratio of living mothers to living daughters, $N_m(t,a_j)/N_d(t,a_j)$,

$$M_1(a) = \int_\alpha^\omega \frac{\bar{l}(x+a_j)}{\bar{l}(x)}e^{-rx}\bar{l}(x)\bar{m}(x)dx = \int_\alpha^\omega e^{-rx}\bar{l}(x+a_j)\bar{m}(x)dx. \tag{6.12}$$

Parent-centric view

Another derivation of the mother-daughter probabilities can be found from what Skaug (2017) call the parent-centric approach. From the parent centric view, the question is given a female, what is the probability that a randomly selected individual is the daughter. The short discussion here ignores the possibility of different sampling probabilities and sampling events to place this within the notation and framework above.

Given both the age a_i of the mother and the age a_j of the daughter, equal sampling probabilities and sampling time, the probability that a randomly selected female is i's daughter can be found as the surviving number of her offspring, relative to the total number of offspring born a_j years ago, i.e.

$$P(\text{i is j's daughter}|a_j,a_i) = \frac{m(a_i-a_j)da_i}{B(t-a_j)}. \tag{6.13}$$

Allowing for heterogeneity (with frailties z_i and z_j),

$$P(\text{i is j's daughter}|a_j,a_i) = \frac{m(a_i-a_j|z_i)da_i l(a_j|z_j)}{B(t-a_j)\bar{l}(a_j)}. \tag{6.14}$$

To replace $m(x|z_i)$ and $l(x|z_j)$ with their marginal values, we need to assume independence between z_i and z_j. Since we are conditioning on the mother being alive, we also need to assume that $m(x|z_i)$ and the survival of the mother is independent. Using (6.8) above and the stable age assumption, $B(t-a_j) = B(t)e^{-ra_j} = N(t)\bar{b}e^{-ra_j}$, and

$$P(\text{i is j's daughter}|a_j,a_i) = \frac{\bar{m}(a_i-a_j)da_i}{N(t)\bar{b}e^{-ra_j}}. \tag{6.15}$$

If the mother's age is unknown, we can find the probability by taking the expectation over the stable age distribution (6.6),

$$\begin{aligned} P(\text{i is j's daughter}|a_j) &= \frac{\int_{a_j}^{\infty} \bar{m}(a_i - a_j)\bar{b}e^{-ra_i}\bar{l}(a_i)da_i}{N(t)\bar{b}e^{-ra_j}} \\ &= \frac{1}{N(t)} \int_0^{\infty} \bar{m}(x)e^{-rx}\bar{l}(x + a_i)dx \\ &= \frac{1}{N(t)} M_1(a_j). \end{aligned} \tag{6.16}$$

6.2 Parent offspring probabilities with both ages unknown

The ages of both parent and offspring may both be unknown for several reasons. In many species, the adults change little after reaching maturity, and determining the age of adults may be difficult or impossible, for example in many whale species. Another case where both ages are unknown is if the samples are collected indirectly. Genetic mark–recapture has been used on non-invasive samples for several species using only animal remains such as hair, faeces or feathers (Lucacs and Burnham 2005). A model where both ages are unknown may be required to extend this non-invasive sampling from self-recapture to kin-recapture, although it may still be possible to separate adults from juveniles.

In the case when both ages are unknown, we can find the probability of an individual having a living mother by integrating over the age distribution of the daughter, $f(a)$,

$$\mathcal{P} = P(\text{Mother alive}) = \int_0^{\infty} M_1(a)f(a)da. \tag{6.17}$$

6.2.1 Constant fecundity

Skaug (2017) gives a proof that if fecundity and population size is constant, then regardless of the shape of the survival function, the probability of having a living mother is 1/2.

$$\begin{aligned} \mathcal{P} = P(\text{Mother alive}) &= \int_0^{\infty} M_1(a)f(a)da \\ &= \int_0^{\infty} \left[\int_0^{\infty} l(a + x)e^{-rx}m(x)dx \right] f(a)da. \end{aligned} \tag{6.18}$$

Assuming constant fecundity, $m(x) = b_0$. The characteristic equation is then $\int_0^{\infty} b_0 \, l(x) \, e^{-rx} \, dx = 1$, and $b_0 = \bar{b}$, the average birth rate in the population, which is also the normalisation constant of the age distribution (6.8). Hence,

$$\begin{aligned} \mathcal{P} &= \bar{b} \int_0^{\infty} \int_0^{\infty} l(a + x)e^{-rx} \, dx f(a) \, da \\ &= \int_0^{\infty} \int_0^{\infty} e^{ra} \, f(a + x) \, dx \, f(a) \, da. \end{aligned} \tag{6.19}$$

Further, assuming that the population size is constant, $r = 0$, and

$$\begin{aligned}\mathcal{P} &= \int_0^\infty \int_0^\infty f(a+x)\,dx\,f(a)\,da \\ &= \int_0^\infty (1 - F(a))f(a)\,da \\ &= \left[F(a) - (F(a))^2/2\right]_0^\infty = 1/2.\end{aligned} \tag{6.20}$$

Comparing (6.19) and (6.20), we see that if the population is increasing, $e^{ra} > 1$, and $\mathcal{P} > 1/2$, and if the population is decreasing, $e^{ra} < 1$, and $\mathcal{P} < 1/2$. Intuitively this reflects the fact that an increasing population is younger, and the age distribution in the population, $f(a)$, is skewed towards zero, and a randomly selected individual is likely to be younger.

6.2.2 Constant mortality

Probability of mother being alive in populations with constant mortality

Theorem 1. *In a population with constant mortality size and stable age distribution, the probability that a randomly selected individual has a a living mother is given by*

$$\mathcal{P} = \frac{r+\mu}{r+2\mu}. \tag{6.21}$$

Proof. The probability of a female of age a_j having a living mother can be found as $M_1(a_j) = \int_0^\infty l(x+a_j)/l(x)g(x)dx$, where $g(x)$ is the age distribution of birthing females in the population. If the population is at stable age distribution, $g(x) = e^{-rx}m(x)l(x)$, and

$$P(M_j \text{ alive}|a_j) = \int_0^\infty e^{-rx}l(x+a_j)m(x)dx. \tag{6.22}$$

See Keyfitz and Caswell (2005, chap 15.2, eq. 15.1.3) and Skaug (2017). If the daughter's age also is unknown, we find the probability of the mother being alive by taking the expectation over the age distribution of the daughters in the population, $f(a_j) = \bar{b}l(a_j)e^{-ra_j}$, where $\bar{b}^{-1} = \int_0^\infty l(x)e^{-rx}$ (Keyfitz and Caswell 2005, eq. 5.1.1, 5.1.2)), and

$$P(M_j \text{ alive}) = \int_0^\infty \int_0^\infty e^{-rx}l(x+a_j)m(x)dx\bar{b}l(a_j)e^{-ra_j}da_j. \tag{6.23}$$

If we assume that the mortality rate, μ is constant, $l(x) = e^{-\mu x}$, and

$$\bar{b} = \left(\int_0^\infty e^{-rx}l(x)\right)^{-1} = \left(\int_0^\infty e^{-x(r+\mu)}\right)^{-1} = r+\mu. \tag{6.24}$$

Combining these,

$$
\begin{aligned}
P(M_j \text{ alive}) &= \int_0^\infty \int_0^\infty e^{-\mu(x+a_j)} m(x) e^{-rx} \bar{b} e^{-\mu a_j} e^{-r a_j} dx da_j && (6.25)\\
&= \bar{b} \int_0^\infty e^{-a_j(2\mu+r)} e^{-x(\mu+r)} m(x) dx da_j && (6.26)\\
&= \bar{b} \int_0^\infty e^{-a_j(2\mu+r)} da_j \int_0^\infty e^{-\mu x} e^{-rx} m(x) dx && (6.27)\\
&= \bar{b} \int_0^\infty e^{-a_j(2\mu+r)} da_j && (6.28)\\
&= \frac{r+\mu}{r+2\mu}. && (6.29)
\end{aligned}
$$

□

In the case where the population is stable ($r = 0$), or high turnover relative to growth ($\mu \gg r$), gives $P(M_j \text{ alive}) = 1/2$, as in the case with constant birth rate.

6.2.3 Constant mortality after maturity

Constant mortality from zygote to old age is a very strong assumption. A more reasonable assumption would be to assume constant mortality only after the onset of maturity (α). If mortality is constant after maturity, the survival of the mother after giving birth to the daughter, $P(M_j \text{ alive}|a_j) = e^{-\mu a_j}$, where μ is the adult mortality rate.

This also follows from the stable age distribution, where $P(M_j \text{ alive}|a_j) = \int_0^\infty e^{-rx} l(x + a_j) m(x) dx$. Since mortality after maturity is constant, $l(x + a_j) = l(x) e^{-\mu a_j}$, and

$$P(M_j \text{ alive}|a_j) = e^{-\mu a_j} \int_0^\infty e^{-rx} m(x) l(x) dx = e^{-\mu a_j}. \tag{6.30}$$

If the daughter's age is unknown, we again find the marginal probability by integrating over the daughher's age distribution,

$$P(M_j \text{ alive}) = \int_0^\infty e^{-\mu x} f(x) dx = \frac{\int_0^\infty e^{-\mu x - rx} l(x) dx}{\int_0^\infty e^{-rx} l(x) dx}, \tag{6.31}$$

which is not very useful, as it requires knowledge of the daughter's mortality curve prior to maturity, which might be rather complex. However, the daughter's age is rarely completely unknown. If an actual individual is sampled (a counter example could be just a genetic sample), the observer should be able determine if the individual has reached the adult stage.

If we know the daughter has reached maturity, we can calculate the conditional probability that the mother is alive. Let α denote the age of maturity (or some age prior to maturity, after which the mortality is considered constant, e.g. yearlings), then

$$P(M_j \text{ alive}|a_j > \alpha) = \int_\alpha^\infty e^{-\mu x} f(x|x > \alpha) dx. \tag{6.32}$$

If $l(\alpha)$ is juvenile survival, and $x > \alpha$, $l(x) = l(\alpha)e^{-\mu(x-\alpha)} = l(\alpha)e^{\mu\alpha}e^{-\mu a}$. The normalisation constant for $f(a_j|a_j > \alpha)$ is then $\int_\alpha^\infty l(\alpha)e^{\mu\alpha}e^{-\mu a}e^{-rx}dx = l(\alpha)e^{-\alpha r}/(r+\mu)$, which gives

$$f(a_j|a_j > \alpha) = \frac{l(a_j)e^{-ra_j}}{\int_\alpha^\infty l(x)e^{-rx}dx} = \frac{l(\alpha)e^{\mu\alpha}e^{-\mu a_j}e^{-ra_j}}{l(\alpha)e^{-r\alpha}(r+\mu)^{-1}} = (r+\mu)e^{-r(a_j-\alpha)-\mu(a_j-\alpha)}, \tag{6.33}$$

and

$$\begin{aligned} P(M_j \text{ alive } |a_j > \alpha) &= \int_\alpha^\infty e^{-\mu x}(r+\mu)e^{-r(x-\alpha)-\mu(x-\alpha)} && (6.34)\\ &= (r+\mu)e^{\alpha(r+\mu)}\int_\alpha^\infty e^{-(2\mu+r)x} && (6.35)\\ &= e^{-\alpha\mu}\frac{\mu+r}{2\mu+r}. && (6.36)\end{aligned}$$

This means that the results for constant mortality above (6.29) is also valid if there is a period of variable mortality prior to maturity, if we only count the daughters that is past this period, and discount with the probability that the adult survives the same period.

If the population is stable, adult mortality constant, and we only observe mature daughters, the probability of the mother being alive is $1/2P(\text{Adult survives 1 maturation period})$.

6.2.4 Varying both fecundity and mortality

The identical results for the edge cases constant fecundity and constant survival suggest that there is a relation between the two results. To see how these two results relate, I will use the covariance function between the fecundity and survival parts to see how the result is dependent on the relation between survival and fecundity. For the covariance term, is should be noted that the only truly random variable here is the age of the mother. The fecundity and the survival term are deterministic given the mother's age, x. The covariance function is here used as a tool to describe how the life history characters vary through the life of an individual.

Theorem 2. *In a population with constant population size and stable age distribution, the probability that a randomly selected individual has a living mother is given by*

$$\mathcal{P} = \frac{1}{2} + COV_X\left[m(x), \int_0^\infty l(a)\frac{l(a+x)}{l(x)}\,da\right]. \tag{6.37}$$

Proof. Starting with eq (6.17), and the assuming constant population size,

$$\begin{aligned} \mathcal{P} &= \int_0^\infty M_1(a)f(a)\,da && (6.38)\\ &= \bar{b}\int_0^\infty\int_0^\infty l(a)m(x)l(x)\frac{l(a+x)}{l(x)}\,da\,dx && (6.39)\\ &= \int_0^\infty\int_0^\infty f(x)\,m(x)\frac{l(a+x)}{l(x)}l(a)\,da\,dx && (6.40)\\ &= \int_0^\infty E_X\left[m(x)l(a)\frac{l(a+x)}{l(x)}\right]\,da. && (6.41)\end{aligned}$$

Separate the E_X term using the definition of covariance,

$$\begin{aligned}\mathcal{P} = \int_0^\infty & E_X\left[m(x)\right]\, E_X\left[l(a)\frac{l(a+x)}{l(x)}\right] \\ & + COV_X\left[m(x), l(a)\frac{l(a+x)}{l(x)}\right] da\end{aligned} \tag{6.42}$$

$$\begin{aligned}\mathcal{P} = E_X[m(x)] & \int_0^\infty E_X\left[l(a)\frac{l(a+x)}{l(x)}\right] da \\ & + \int_0^\infty COV_X\left[m(x), l(a)\frac{l(a+x)}{l(x)}\right] da.\end{aligned} \tag{6.43}$$

Consider the terms separately. The expectation of the fecundity is found from the stable population size assumption,

$$E_X[m(x)] = \int_0^\infty f(x)m(x)\,dx = \bar{b}\int_0^\infty l(x)m(x)\,dx = \bar{b}R_0 = \bar{b}. \tag{6.44}$$

The second expectation follows from (6.20),

$$\begin{aligned}\int_0^\infty E_X\left[l(a)\frac{l(a+x)}{l(x)}\right] da &= \int_0^\infty\int_0^\infty l(a)\frac{l(a+x)}{l(x)}f(x)\,da\,dx && (6.45)\\ &= \bar{b}\int_0^\infty\int_0^\infty l(a+x)l(a)\,da\,dx && (6.46)\\ &= \frac{1}{\bar{b}}\int_0^\infty\int_0^\infty f(a+x)f(a)\,da\,dx && (6.47)\\ &= \frac{1}{2\bar{b}}. && (6.48)\end{aligned}$$

Which gives the following expression for $\mathcal{P}$

$$\begin{aligned}\mathcal{P} &= \frac{1}{2} + \int_0^\infty COV_X\left[m(x), l(a)\frac{l(a+x)}{l(x)}\right] da && (6.49)\\ &= \frac{1}{2} + COV_X\left[m(x), \int_0^\infty l(a)\frac{l(a+x)}{l(x)}\,da\right]. && (6.50)\end{aligned}$$

□

The edge cases, constant mortality and constant fecundity follows directly from the properties of the covariance function. If fecundity, $m(x)$, is constant, the fecundity term of the covariance expression is independent of x. If mortality is constant, $l(a+x)/l(x) = l(a)$, and the survival term of the covariance is constant. In both cases, the covariance term disappears, and $\mathcal{P} = 1/2$ as expected.

If we assume that fecundity and mortality are monotonous functions of age, the sign of the covariance term can be found, using that both sides are functions of the same random variable.

For the survival term, let

$$\Lambda(x) = l(a)\frac{l(a+x)}{l(x)} = e^{-M(a+x)-M(a)+M(x)}, \tag{6.51}$$

where $M(x) = \int_0^x \mu(t)dt$, then

$$\frac{\partial \Lambda}{\partial x} = (\mu(x) - \mu(a+x))\Lambda(x). \tag{6.52}$$

I.e. $\Lambda(x)$ is a decreasing function if the hazard rate, $\mu(x)$, is an increasing function of age, and decreasing if hazard is increasing with age. It follows that if both mortality rate and fecundity increase with age, or both decrease, the sign of the covariance term is negative, and $\mathcal{P} < 1/2$. Likewise, if the age derivative of fecundity and the mortality rate are of opposite sign, $\mathcal{P} > 1/2$. The survival term, $\int_0^\infty l(a)l(a+x)/l(x)\,da$, is the conditional survival of the mother averaged over the age distribution of the daughters. The separation of fecundity and survival is dependent on the assumption of a stable population. If the population is growing or shrinking, the constant term will differ from 1/2, as discussed in section 6.2.1.

6.2.5 Discussion

The identical result in the two edge cases, constant fecundity and constant mortality is interesting, as these two cases can describe very different life histories, and suggests that $\mathcal{P} = 1/2$ might be used as an approximation when little is known about the population parameters. The covariance formulation of $\mathcal{P}$ in Theorem 2 can give some guidance on when the approximation is appropriate, and the direction of bias if the assumptions are violated.

The biggest issue with constant fecundity or mortality models is the assumption that the rates are constant from the moment they are born. For a species where the animals change little after maturity, it may be a valid assumption that the mortality is near constant after maturity, and the results in section 6.2.3 may be used in a CKMR model. Since only the adults are used in (6.36), a CKMR model using this relationship would only have to assume that the adults are a random sample.

6.3 Survival dependencies between mother and daughter

A major assumption in section 6.1 is that the survival of mothers and daughters are independent. This assumption may be violated in several ways. An obvious example is heritable differences, or different genetic variants of the same species. However, dependencies in survival might also arise from non-genetic effects. For instance, an offspring might share a less profitable habitat with its mother, or a low status female may produce offspring with lower status due to parental care or other maternal effects.

6.3.1 Distribution of age and frailty in mothers

A common approach to unobserved heterogeneity affecting survival is frailty models. In frailty models, survival is dependent on some unobserved random variable Z, assigned at birth (Vaupel and Missov 2014).

To study how heterogeneity affects the mother-daughter probabilities, we will need the distribution of age and phenotype of birthing females in the population. The purpose of the following is to see how the frailty variable affects this distribution. Let $g(x, z)$ denote the fraction of the births by mothers of age x and phenotype z, and $B(t)$ the total number of births in the population at a point in time. Of the $B(t-x)$ births x years ago, $\pi(z)$ has phenotype z, a fraction $l(x|z)$ survives to age x, and gives birth to $m(x|z)$ offspring.

$$g(x,z) = B(t-x)\pi(z)l(x|z)m(x|z)/B(t). \tag{6.53}$$

A population with stable vital rates will converge to a stable age distribution, where the number of births in the population increases at a rate r, so that the ratio $B(t)/B(t-x) = e^{rx}$. If we assume the population follows a stable age distribution, the bivariate density can be written as

$$g(x,z) = \pi(z)l(x|z)m(x|z)e^{-rx}. \tag{6.54}$$

To find the marginal distribution of x, $\bar{g}_X(x)$ we have to integrate out z:

$$\bar{g}_x(x) = \int_0^\infty e^{-rx}\pi(z)l(x|z)m(x|z)dz. \tag{6.55}$$

In the general case where fecundity can also be a function of z, the marginal distributions $l(x)$ and $m(x)$ will not determine the marginal motherhood distribution. For convenience, define the net maternity function $k(x|z) = m(x|z)l(x|z)$, and let $g(x) = e^{-rx}k(x)$. Otherwise If $m(x)$ is independent of z, $\bar{l}(x) = \int_0^\infty \pi(z)l(x|z)dz$, and $g(x)$ can be written as the usual

$$\bar{g}_X(x) = e^{-rx}\bar{l}(x)m(x). \tag{6.56}$$

The marginal distribution of z for the mothers for all births in the population, $\bar{g}_Z(z)$,

$$\bar{g}_Z(z) = \pi(z)\int_0^\infty e^{-rx}l(x|z)m(x|z)dx = \pi(z)R(z), \tag{6.57}$$

where $R(z) = \int_0^\infty e^{-rx}l(x|z)m(x)dx$ can be regarded as the expected reproductive contribution of individuals of type z. In the case of a stationary population this is simply the expected number of offspring of a female of type z. The conditional distributions (assuming $m(x)$ is independent of z) can be found from the above:

$$g(z|x) = \frac{g(x,z)}{\bar{g}_Z(z)} = \pi(z)\frac{l(x|z)}{\bar{l}(x)}, \tag{6.58}$$

which is just the distribution of frailty at age x (Vaupel and Missov 2014, eq. 8). The conditional distribution of age given frailty can similarly be found as

$$g(x|z) = \frac{g(x,z)}{\bar{g}_Z(z)} = \frac{e^{-rx}l(x|z)m(x)}{\int_0^\infty e^{-rx}l(x|z)m(x)dx} = \frac{e^{-rx}l(x|z)m(x)}{R(z)}. \tag{6.59}$$

6.3.2 Probability of living mother

To account for heterogeneity that affects both the mother and the daughter, we need to account for that the mortality of the mother and daughter are not independent, i.e., $P(M_j \text{ alive}|a_j, j \text{ alive}) = P(M_j \text{ alive}, j \text{ alive}|a_j)/l(a_j)$.

By counting method

Following the counting method approach described in section 6.1.6, but relaxing the independence assumption, we assume that the mother and daughter are assigned a latent variable z_i and z_j at birth, which affects survival and take expectations,

$$\begin{aligned} M_1(a_j) &= P(M_j \text{ alive}|a_j, j \text{ alive}) \\ &= \frac{E_{Z_i,Z_j}\int_\alpha^\infty B(t-a_j-x)l(x|z_i)m(x)\frac{l(x+a_j|z_i)}{l(x|z_i)}l(a_j|z_j)dx}{B(t-a_j)\bar{l}(a_j)}. \end{aligned} \tag{6.60}$$

Using stable age assumption, the total number of births grows by a rate r, so that $B(t-a-x)/B(t-a) = e^{-rx}$. Inserting this and simplifying,

$$M_1(a_j) = 1/\bar{l}(a_j)\int_\alpha^\infty e^{-rx}m(x)E[l(x+a|z_i)l(a|z_j)]dx. \tag{6.61}$$

From the definition of covariance,

$$E[l(x+a|z_i)l(a|z_j)] = E[l(x+a|z_i)]E[l(a|z_j)] + cov(l(x+a|z_i), l(a|z_j)), \tag{6.62}$$

and setting $E_{Z_i}[l(x+a|z_i)] = \bar{l}(a)$,

$$M_1(a_j) = \int_\alpha^\infty e^{-rx}m(x)\bar{l}(x+a_j)dx + \int_\alpha^\infty \frac{m(x)}{\bar{l}(a_j)}e^{-rx}cov(l(x+a|z_i), l(a_j|z_j))dx. \tag{6.63}$$

The first term gives the probability of living mother when there is no correlation between the mortality of the mothter and daughter.

By conditional probability

The same result can be derived inn a somewhat less straightforward way from the individual point of view by using conditional probabilities and the results from section 6.3.1. Given a female of age a_j, the conditional distribution of frailty, z_j can be found as $\pi(z_j)l(a_j|z_j)/\bar{l}(a_j)$ (Vaupel and Missov 2014, eq. 8). Define $h(z_i, z_j)$ as the joint distribution of frailty in mother and daughter, $h_m(z_i|z_j)$ as the conditional distribution of frailty in mothers given daughters frailty, and $h_d(z_j|z_i)$ as the conditional distribution of frailty in daughters, given the mother's frailty. If the mother's frailty z_i is known, the conditional age distribution of living mothers of daughters born a_j year ago can be found using (6.59) above as

$$g(x|z_i)\frac{l(x+a_j|z_i)}{l(x|z_i)} = \frac{l(x+a_j|z_i)m(x)e^{-rx}}{R(z_i)}. \tag{6.64}$$

Combining this, and integrating over the frailties of both mother and daughter,

$$M_1(a_j) = \int_0^\infty \int_0^\infty \int_0^\infty \frac{l(x+a_j|z_i)m(x)e^{-rx}}{R(z_i)}\,dx\, h_m(z_i|z_j)dz_i\,\frac{\pi(z_j)l(a_j|z_j)}{\bar{l}(a_j)}dz_j. \tag{6.65}$$

Using Bayes' equation to write the heritability term in terms of the daughter inheriting the mother we have that

$$h_m(z_i|z_j) = \frac{h_d(z_j|z_i)\bar{g}_Z(z_i)}{\int_0^\infty h_d(z_j|z)\bar{g}_Z(z)dz}. \tag{6.66}$$

For the model to be stationary, the marginal distribution of frailty in the newborn daughters must be the same as the initial distribution of frailties in the mothers, i.e.

$$\pi(z_j) = \int_0^\infty h_d(z_j|z)\bar{g}_Z(z)dz. \tag{6.67}$$

Assuming stationarity, we have that

$$h_m(z_i|z_j) = \frac{h_d(z_j|z_i)\bar{g}_Z(z_i)}{\pi(z_j)} = \frac{h_d(z_j|z_i)\pi(z_i)R(z_i)}{\pi(z_j)}. \tag{6.68}$$

Inserting this into (6.65),

$$\begin{aligned} M_1(a_j) &= \frac{1}{\bar{l}(a_j)}\int_0^\infty \int_0^\infty \int_0^\infty l(x+a_j|z_i)l(a_j|z_j)m(x)e^{-rx}h_d(z_j|z_i)\pi(z_i)\,dz_i\,dz_j\,dx && (6.69)\\ &= \frac{1}{\bar{l}(a_j)}\int_\alpha^\infty e^{-rx}m(x)E[l(x+a_j|z_i)l(a|z_j)]dx. && (6.70)\end{aligned}$$

If frailty is independent between mother and daughter, $h_d(z_j|z_i)\pi(z_i) = \pi(z_j)\pi(z_i)$, and (6.69) simplifies to (6.12), as expected.

6.3.3 Discussion

The objective of this section was to study how dependencies in survival between mother and daughter affect the probability of having a living mother. The basic idea was to use the stable age distribution and frailty concepts to model the within population variability.

Moving beyond the basic model setup to model the dependencies between mother and daughter in a consistent way takes some care, both in formulation and interpretation of the model. If the dependency in survival between parent and offspring is caused by simple genetic inheritance, the genetic composition will change from generation to generation by the process of natural selection. This process is well understood (Charlesworth 1994, e.g.), the natural selection process will maximise the fitness, i.e. the r in the Euler-Lotka equation. In the model here, the less frail individual will reproduce less, and if the frailty is directly heritable, the distribution of frailty in the newborn will be skewed toward lower frailty in the newborn cohort compared to the parent population.

To create a statistical model where there is persistent variability in survival, but dependent within parent-offspring pairs, we need a model to be stationary across generations, i.e. the distribution of frailty in the newborns should be the same (6.67).

A possible way to model the mother-daughter dependencies is using a shared frailty approach, where the total frailty of an offspring is a combination of the mothers frailty and its own frailty. As an example, let $l_j(a|z_i, z_j) = exp\left(-\int_0^a \mu_1(x, z_j)dx - \int_0^a \mu_1(x, z_i)dx\right) = l_{j0}(x|z_i)l_{j1}(x|z_j)$, where z_j is the frailty peculiar to j, and z_i is the shared component of the frailty. The added own frailty of the daughter would ensure that the variation in frailty is not eliminated by the selective process. Viewing the dependencies this way, as a stochastic process from mother to daughter, has its own difficulties, both in proving that it converges to a limiting distribution for frailty at birth, and in sampling from this distribution. After the initial consideration of this and some similar models where the frailty was modelled as a stochastic process, I decided not to put more effort into the development of this model. This was based both on the mathematical intractability and the lack of a clear biological interpretation of the model.

Part of the problem with the approach discussed above, is that the dependencies also induces subtle changes to the population age structure. If the dependencies changes the initial distribution of frailty at birth, $\pi(z)$, this will also alter the marginal age composition in the population, $\bar{l}(a)$, and therefore affect the whole population equation, including the population growth rate, r. An attempt to model various levels of dependency within the mother-daughter probabilities would also include the effects the dependency model have on the age composition. A possible way to avoid the coupling between the population composition and the mother-daughter dependencies could be to use an approach based on copulas. Using copulas, it is possible to model the marginal survival probabilities and the dependencies separately, while in a shared frailty model the marginal survival probabilities are a function of the frailty distribution (Goethals et al. 2008).

Even if the frailty of parent and offspring is strongly correlated, knowing that an offspring is alive might only be weakly informative about whether the parent is alive. Conditional on the frailty, the survival of parent and offspring is independent in a correlated or shared frailty model. Even if there is high correlation between the frailties, the frailty may explain only a small part of the variation in survival. Vaupel (1988) discusses a model of inherited frailty, and notes that even if there is substantial heterogeneity in frailty and a high correlation between the frailties of parent and children, there is often a weak correlation between the life spans of parents and children.

Much of the complexity of this approach stems from the distribution of frailties in birthing females and the connection between this and the age distribution of the females. For future research into correlated survival in a CKMR setting, it might be a more tractable starting point to consider a model conditional on the mother's age, closer to the expected relative reproductive output model where the first assumption (1.9) is violated. If only a single parent cohort is considered, it might be possible to assume a standard distribution for the frailty in the parent cohort, and ignore the feedback on population structure.

Final remarks

It is customary to end with the conclusion that further research is needed. In a field as young as Close-Kin Mark-Recapture, this is especially true. The interplay between statistics, population biology and genetics offers a wide range of unexplored problems. It has been enjoyable doing my thesis in a field where there still is so much uncharted ground.

The chapters of this thesis have covered different aspects of CKMR methods. A unifying theme through the various chapters is the lifetime changes and variation in fecundity and heterogeneity. The dependency between the reproductive output and recapture probabilities is a key feature of CKMR models. In the parent-offspring CKMR in chapters 2 and 3 the age and age specific fecundities are considered known, and the heterogeneity in reproduction can be regarded as an observed heterogeneity. In the half sibling chapters 4 and 5, the age and fecundities of the parents are unknown, and an unobserved heterogeneity.

Parent-offspring To expand the estimator in chapters 2 and 3 to other parameters than population size maximum likelihood methods are needed. Chapter 3 suggests that by viewing the number of marks in the population as a random quantity, it is possible to construct a conditional likelihood for the number of recaptured offspring given the parent's observed covariates. This approach should make it possible to avoid several of the problems with the pseudo likelihood approach which arise in small populations and non-sparse sampling. Further research is needed to expand on this.

Half siblings The half sibling analysis in chapter 4 only briefly touch upon the demographic questions which need to be addressed for a half sibling analysis. The indirect nature of the half sibling analysis makes it more challenging than parent-offspring analysis. Half siblings can be used to estimate population parameters not observable with parent offspring CKMR. It is for example not possible to estimate mortality with parent-offspring CKMR if sampling is lethal. For the constant mortality case, the simple model in (1.17) can be used to estimate mortality. However, in organisms such as fish, where survival and fecundities change through the life, the half sibling probabilities depend on several confounded parameters, as noted in section 4.6. Incorporating half siblings in a CKMR demographic model beyond the constant mortality case is non-trivial, and an area that need more research.

A question left unanswered in the same cohort half sibling chapters, is what information about the heterogeneity in reproduction can be found from the size of the observed sibling clusters. If sampling is non-sparse, a sample may be expected to contain a number of sibling groups, not only occasional pairs. In the discussion in 4.6, I give some reasons why this may not be sufficient to describe the heterogeneity. It may, however, still be a question worth pursuing further.

Demographic models Continuous time demographic models have the advantage that it help us arrive at closed form solutions, as for example the results in section 6.2, for the case when age information is unknown. They are, however, not easily extended to population models where the dynamics is not well described by age. In many cases age gives an incomplete, or even misleading description of population structure. The vital rates of individuals will for example in many cases depend on the size rather than age. If growth varies among individuals, a model based on size may better capture the dynamics of the population. The framework of matrix population models (Caswell 2001) extends the concept of stable age structure to that of stable stage (or state) distributions, and allow for a greater flexibility for the construction of population models. Describing within population heterogeneity in survival using the continuous demographic framework proved difficult. A state-based model using the population projection matrices may offer more flexible way of describing heterogenous populations than what is possible using the age only continuous framework.

Reproductive success The relationship between reproductive success and individual demographic parameters as size and age may be considered a nuisance in the CKMR setting. From an ecological point of view these are very interesting parameters, which have important implications for both population dynamics, population viability and evolutionary dynamics. Reproductive success depends on both behaviour and physical parameters, and estimates based on body size alone, or even controlled experiments may differ considerably from the actual reproductive success in a natural environment where the animals are subject to competition, predation and other ecological processes.

As the CKMR methods are applied to more populations, they can also contribute to a better understanding of the importance of life history variables like size in natural populations, as exemplified by the southern bluefin tuna study in Bravington et al. (2016a).

Appendix A

Discrete probability distributions

This appendix gives a short overview of the probability distributions for count data referred to in the main text.

Binomial distribution

If we draw n individuals with replacement from a population of size N with K marks, the probability of drawing k marked individuals are described by the binomial distribution,

$$P(X = k|n, K, N) = \binom{n}{k} \left(\frac{K}{N}\right)^k \left(1 - \frac{K}{N}\right)^{n-k}, \tag{A.1}$$

$$E(X) = n\frac{K}{N}, \tag{A.2}$$

$$VAR(X) = n\frac{K}{N}\frac{(N-K)}{N}. \tag{A.3}$$

If K/N is small, the binomial distribution can be approximated by the Poisson distribution (law of rare events).

Hypergeometric distribution

If we draw n individuals without replacement from a population of size N with K marks, the probability of drawing k marked individuals are described by the hypergeometric distribution,

$$P(X = k|n, K, N) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \tag{A.4}$$

$$E(X) = n\frac{K}{N}, \tag{A.5}$$

$$VAR(X) = n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}. \tag{A.6}$$

If $N \gg n$, the probability of drawing the same individual is low, and the hypergeometric distribution can be approximated by the binomial distribution. The variance of the hypergeometric distribution is smaller than the variance of the binomial by the factor $N - n/N - 1$, which is often referred to as the finite population correction factor.

Poisson distribution

The Poisson distribution arise in two situations discussed here. One is as the limiting distribution of the binomial or hypergeometric distributions as population size tends to infinity. The other is as number of counts in a Poisson process, e.g. if births occur randomly at a fixed rate in a population, the number of births for an individual in a period of time has the Poisson distribution.

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{A.7}$$

$$E(X) = \lambda, \tag{A.8}$$

$$VAR(X) = \lambda. \tag{A.9}$$

The Poisson distribution is a single parameter distribution with mean equal to the variance. When the term overdispersed is used here, it should be understood as overdispersed compared to the Poisson distribution, i.e. any count distribution with variance higher than the mean.

Negative binomial distribution

The negative binomial distribution is a commonly used distribution to model overdispersed count data. Parametrised in terms of the mean λ and dispersion parameter ϕ the negative binomial distribution has

$$Pr(X = k|\phi, \lambda) = \frac{\Gamma(\phi + k)}{k!\,\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda}\right)^r \left(\frac{\lambda}{\phi + \lambda}\right)^k, \tag{A.10}$$

$$E(X) = \lambda, \tag{A.11}$$

$$VAR(X) = \lambda + \frac{\lambda^2}{\phi}. \tag{A.12}$$

Rearranging the variance gives

$$\phi^{-1} = VAR(X)/\lambda^2 - \lambda^{-1} = CV(X)^2 - 1/\lambda. \tag{A.13}$$

The negative binomial is asymptotically equal to the Poisson distribution as $\phi^{-1} \to 0$.

Appendix B

Bibliography

Akita, T., 2018. Statistical test for detecting overdispersion in offspring number based on kinship information. Population Ecology 60 (4), 297–308.

Akita, T., 2019. Nearly unbiased estimator of adult population size based on within-cohort half-sibling pairs incorporating flexible reproductive variation. bioRxiv (ii), 1–27.
URL https://www.biorxiv.org/content/early/2018/09/20/422659

Bailey, N. T. J., 1951. On Estimating the Size of Mobile Populations from Recapture Data. Biometrika 38 (3), 293–306.

Bravington, M. V., Eveson, J. P., Grewe, P. M., Davies, C. R., 2017. SBT Close-Kin Mark-Recapture with Parent-Offspring and Half-Sibling Pairs : update on genotyping, kin-finding and model development (Scientific Committee Report CCSBT-ESC/1709/12). Tech. rep., CSIRO, Hobart, Australia.

Bravington, M. V., Grewe, P. M., Davies, C. R., 2016a. Absolute abundance of southern bluefin tuna estimated by close-kin mark-recapture. Nature Communications 7, 1–8.
URL http://dx.doi.org/10.1038/ncomms13162

Bravington, M. V., Skaug, H. J., Anderson, E. C., 2016b. Close-Kin Mark-Recapture. Statistical Science 31 (2), 259–274.

Caswell, H., 2001. Matrix Population Models: Construction, Analysis, and Interpretation, 2nd Edition. Sinauer Associates.

Chao, A., Pan, H. Y., Chiang, S. C., 2008. The Petersen - Lincoln estimator and its extension to estimate the size of a shared population. Biometrical Journal 50 (6), 957–970.

Chapman, D. G. ., 1951. Some properties of the hypergeometric distribution with applications to zoological sample censuses. Univ. California Publ. Stat. 1, 131–160.
URL https://catalog.hathitrust.org/Record/005761134

Chapman, D. G. ., 1952. Inverse , Multiple and Sequential Sample Censuses. Biometrics 8 (4), 286–306.
URL https://www.jstor.org/stable/3001864

Charlesworth, B., 1994. Evolution in Age-Structured Populations, second ed Edition. Vol. 1 of Cambridge Studies in Mathematical Biology. Cambridge University Press, Cambridge.

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Sy, 1695.
URL `http://igraph.org`

Fleming, I. A., 2011. Pattern and variability in the breeding system of Atlantic salmon (Salmo salar), with comparisons to other salmonids. Canadian Journal of Fisheries and Aquatic Sciences 55 (S1), 59–76.

Forseth, T., Barlaup, B. T., Finstad, B., Fiske, P., Gjøsæter, H., Falkegård, M., Hindar, A., Mo, T. A., Rikardsen, A. H., Thorstad, E. B., Vøllestad, L. A., Wennevik, V., 2017. The major threats to Atlantic salmon in Norway. ICES Journal of Marine Science 74 (6), 1496–1513.

Goethals, K., Janssen, P., Duchateau, L., sep 2008. Frailty models and copulas: similarities and differences. Journal of Applied Statistics 35 (9), 1071–1079.
URL `https://www.tandfonline.com/doi/full/10.1080/02664760802271389`

Harvey, A. C., Tang, Y., Wennevik, V., Skaala, Ø., Glover, K. A., sep 2017a. Data from: Timing is everything: Fishing-season placement may represent the most important angling-induced evolutionary pressure on Atlantic salmon populations.
URL `https://datadryad.org/resource/doi:10.5061/dryad.r258f9q/1`

Harvey, A. C., Tang, Y., Wennevik, V., Skaala, Ø., Glover, K. A., 2017b. Timing is everything: Fishing-season placement may represent the most important angling-induced evolutionary pressure on Atlantic salmon populations. Ecology and Evolution 7 (18), 7490–7502.

Hillary, R. M., Bravington, M. V., Patterson, T. A., Grewe, P., Bradford, R., Feutry, P., Gunasekera, R., Peddemors, V., Werry, J., Francis, M. P., Duffy, C. A., Bruce, B. D., 2018. Genetic relatedness reveals total population size of white sharks in eastern Australia and New Zealand. Scientific Reports 8 (1), 1–9.

Jones, O. R., Wang, J., 2010. COLONY: A program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10 (3), 551–555.

Keyfitz, N., 1968. Introduction to the Mathematics of Populations. Addison-Wesley.

Keyfitz, N., Caswell, H., 2005. Applied Mathematical Demography. Statistics for Biology and Health. Springer-Verlag, New York.
URL `http://link.springer.com/10.1007/b139042`

Lucacs, P. M., Burnham, K. P., nov 2005. Review of capture-recapture methods applicable to noninvasive genetic sampling. Molecular Ecology 14 (13), 3909–3919.
URL `http://doi.wiley.com/10.1111/j.1365-294X.2005.02717.x`

Pew, J., Muir, P. H., Wang, J., Frasier, T. R., 2015. related: An R package for analysing pairwise relatedness from codominant molecular markers. Molecular Ecology Resources 15 (3), 557–561.

R Core Team, 2018. R: A Language and Environment for Statistical Computing.
URL https://www.r-project.org/

Rawding, D. J., Sharpe, C. S., Blankenship, S. M., 2014. Genetic-Based Estimates of Adult Chinook Salmon Spawner Abundance from Carcass Surveys and Juvenile Out-Migrant Traps. Transactions of the American Fisheries Society 143 (1), 55–67.

Ruzzante, D. E., McCracken, G. R., Førland, B., MacMillan, J., Notte, D., Buhariwalla, C., Flemming, J. M., Skaug, H., 2019. Validation of close-kin mark-recapture (CKMR) methods for estimating population abundance. In revision.

Ruzzante, D. E., McCracken, G. R., Parmelee, S., Hill, K., Corrigan, A., MacMillan, J., Walde, S. J., 2016. Effective number of breeders, effective population size and their relationship with census size in an iteroparous species, Salvelinus fontinalis. Proceedings of the Royal Society B: Biological Sciences 283 (1823).

Schwartz, M. K., Allendorf, F. W., Luikart, G., Tallmon, D. A., Ryman, N., 2010. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. Conservation Genetics 11 (2), 355–373.

Seber, G. A. F., 1982. The estimation of animal abundance and related parameters, 2nd Edition. Edward Arnold: London., London.

Skaala, Ø., Knutar, S., Østebø, B. I., Holmedal, T.-E., Skilbrei, O. T., Sami Madhun, A., Barlaup, B., Urdal, K., 2015. Erfaringar med Resistance Board Weir-fangstsystemet i Etnevassdraget 2013-2014, 1–22.

Skaug, H. J., 2001. Allele-Sharing Met hods for Estimation of Population Size. Biometrics 57 (September), 750–756.

Skaug, H. J., 2017. The parent–offspring probability when sampling age-structured populations. Theoretical Population Biology 118, 20–26.
URL http://linkinghub.elsevier.com/retrieve/pii/S0040580917300138

Skaug, H. J., Bérubé, M., Palsbøll, P. J., 2010. Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate. Molecular Ecology Resources 10 (4), 693–700.

Städele, V., Vigilant, L., 2016. Strategies for determining kinship in wild populations using genetic data. Ecology and Evolution 6 (17), 6107–6120.

Vaupel, J. W., may 1988. Inherited Frailty and Longevity. Demography 25 (2), 277.
URL http://link.springer.com/10.2307/2061294

Vaupel, J. W., Missov, T. I., 2014. Unobserved population heterogeneity: A review of formal relationships. Demographic Research 31 (1), 659–686.

Wang, J., 2009. A new method for estimating effective population sizes from a single sample of multilocus genotypes. Molecular Ecology 18 (10), 2148–2164.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL `http://ggplot2.org`