New methods for estimating population size based on close-kin genetics and extensions

Robin Aldridge-Sutton

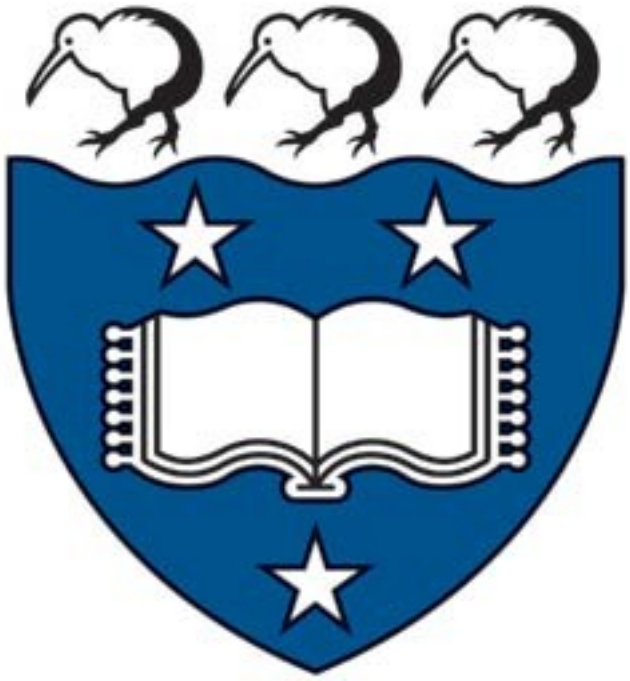
Bachelor of Science (Honours)
Department of Statistics
The University of Auckland
New Zealand

October 2019

Contents

Chapter 1

Acknowledgements



My supervisors are Professor Rachel Fewster, in the Department of Statistics, and Rutherford Discovery Senior Research Fellow Emma Carroll, in the School of Biological Sciences, both at the University of Auckland. They have been generous and patient with me, letting meetings go on as long as necessary, replying to emails quickly and in detail, and gently correcting multiple drafts of this dissertation. All this under sometimes very difficult circumstances. They formulated the plan for this work, and they helped me figure out how to implement it. Working on this project with each of you has been the highlight of the year for me, I have been lucky to have you, and I am grateful to you both.

Chapter 2

Abstract

Close-kin genetics is an emerging approach to estimating population size and demographics, based on family relationships elicited from samples of genotyped animals. The field is still in its infancy but is generating considerable interest, partly due to recent advances in genomics providing high-resolution genotype datasets. Some flagship studies have already been completed, including estimates of population demographics for white sharks in Australia and New Zealand.

We are interested in estimating the size and demographics of the New Zealand southern right whale population (Tohorā), which is now recovering from a dramatic decline after an estimated 35,000 whales were slaughtered during 19th century whaling. We outline several improvements to existing R code for simulating and estimating population trajectories using the classical capture-recapture model 'POPAN', resulting in an increase in computational speed of between four- and five-fold. We then describe new R code for simulating populations under more realistic scenarios, in respect of breeding behaviour, abundance trajectories, and genetic inheritance. We present a Shiny interface for investigating relevant characteristics of the simulated populations.

We then present new ideas for extending close-kin genetics to a wider methodology for estimating population abundance and demographics. Instead of requiring the analyst to predetermine kinship between each pair of animals using a threshold approach, and delete any pairs over which there is doubt, the new method is based on a pseudo-likelihood which incorporates multiple levels of kinship together with kinship uncertainty. The pseudo-likelihood is based on a saddlepoint approximation to the probability density function of a measure of kinship strength between each pair of DNA samples. We present preliminary estimation results based on our simulated right whale population to show the method produces usefully precise information on abundance, survival and growth rates.

Chapter 3

Introduction

Close-kin genetics is an emerging approach to estimating population size and demographics, based on family relationships elicited from samples of genotyped animals (Bravington et al., 2016; Hillary et al., 2018). Before tackling the problem of implementing close-kin models we began by writing R code to simulate data and fit the classical capture-recapture 'POPAN' models, which represent a simpler approach to estimating population size and demographics. We integrated our code into a previously developed but unreleased application, and achieved a 5x improvement in data simulation speed, and a 4x improvement in model-fitting speed. In section 4 we describe this work both for its value in itself, and as an overview of the issues that are discussed in more detail for close-kin methods in the rest of this dissertation.

We are interested in applying close-kin genetics to the New Zealand southern right whale population (*Eubalaena australis*), which is now recovering from a dramatic decline when an estimated 35,000 whales were slaughtered during 19th century whaling (Jackson et al., 2016; Carroll et al., 2013). This is a global pattern and recovering right whale populations have been the subject of long-term monitoring for over four decades, providing useful information for parameterizing simulations (Harcourt et al., 2019; IWC, 2001). In section 5 we present R code for individual-level simulation of this population under realistic scenarios for population and catch trajectories, life history traits, and genetic inheritance. We believe that such detailed simulations are as yet uncommon, and will be useful in their own right as tools to investigate population genetics, as well as to produce data for developing new genetic modelling techniques. We also present a Shiny interface for viewing and validating the resulting datasets.

Close-kin genetics involves calculating a test statistic between each pair of genotypes, namely a pseudo log likelihood ratio, for comparing the hypotheses that the pair has a specified closer kinship versus a specified less-close kinship. This pseudo log likelihood ratio has been dubbed the 'PLOD' in previous literature for an easily-pronounced abbreviation. For example, a PLOD might compare the hypothesis that the pair is a parent-offspring pair with the hypothesis that the pair is unrelated. However, such a measure need not be restricted to investigating parent-offspring relationships, but can be used more generically as a measure

of kinship strength, creating a distribution of PLODs among all pairs in a population such that closely-related pairs score higher values and unrelated pairs score lower values. We hope to develop a new method for close-kin estimation that incorporates multiple levels of kinship together with kinship uncertainty, which we approach using a pseudo-likelihood of the PLODs observed among the pairs in our data. We calculate the pseudo-likelihood function by partitioning over true kinship level, using saddlepoint approximations to find the density of PLODs within each kinship level, and combining them into the overall density using the probabilities of the corresponding kinships given our parameters of interest.

In section 6 we describe how PLODs are calculated and show how their distribution depends in a systematic way on the kinships among the genotyped animals. We then show how to approximate the distribution of PLODs for each particular true kinship level using the saddlepoint approximation. We present the results of our application of this method to data generated using our simulation framework.

In section 7 we derive expressions for three kinpair probabilities: parent-offspring pairs among genotypes from one sample, self-pairs among genotypes from different samples, and parent-offspring pairs among genotypes from samples that are separated by a period which is less than the age of maturity. We show that the numbers of each of these kinpairs occurring in datasets generated using our simulation framework are distributed around those predicted by these expressions.

In section 8 we present a preliminary method which forms the pseudo-likelihood using saddlepoint approximations and kinpair probabilities for just these three kinships. We show that it can be maximised over our parameters of interest using numerical methods. We use our simulation framework to show that even this preliminary method can produce usefully precise estimates of abundance, survival and population growth rate from realistic sample sizes and numbers of samples.

A selection of the most important R code described in this dissertation is available to view at https://github.com/rasutt/Dissertation. It is intended as proof of work, not for reproduction of results or application to new problems.

Chapter 4

Fast POPAN data simulation and model-fitting

POPAN models represent an approach to estimating population size and demographics that is simpler than close-kin genetics in that it is based only on the recapture of uniquely identifiable individuals, not their kin. Before tackling the problem of implementing close-kin models we began by writing R code to simulate data and fit POPAN models. We integrated our code into a previously developed but unreleased application, and achieved a 5x improvement in data simulation speed, and a 4x improvement in model-fitting speed. We describe this work here both for its value in itself, and as an overview of the issues that are discussed in more detail for close-kin methods in the rest of this dissertation.

4.1 POPAN models

POPAN models (Schwarz and Arnason, 1996) are used to estimate the size of a population N, the proportion of animals first entering the population at each survey occasion p_{ent}, and annual survival rate ϕ. They use a likelihood based on capture-recapture data from repeated surveys that capture and identify animals. The likelihood is given by:

$$L(N_s, \theta) = \binom{N_s}{n} p_\theta^n (1 - p_\theta)^{N_s - n} \prod_{i=1}^{n} \left\{ \frac{P(x_i; \theta)}{p_\theta} \right\}, \tag{4.1}$$

where:

- N_s is the superpopulation parameter, representing the number of animals that are ever exposed to capture,
- θ is the vector of other parameters including survival and entry probabilities and capture probabilities,

- n is the total number of animals that are captured during the study,
- p_θ is the probability for each animal in the superpopulation that it is caught at least once, and
- x_i is the capture history for the i-th animal that has been caught at least once.

It is helpful to understand this as a binomial detectability model (Fewster and Jupp, 2009), where detection refers to an animal being caught at least once. The first part of the likelihood is the probability of detecting n animals. This is a binomial random variable parameterised by the number of animals ever exposed to capture and the probability for each one that it is detected. The second part is the product of the conditional probabilities of the capture histories given that the animals are detected.

Capture histories are represented as strings of ones and zeros representing captures and non-captures in each survey. The probabilities of capture histories are found using recurrence relations for leading and trailing zeros to accommodate cases in which the animals are not yet born or have died, as well as cases in which they are alive but not captured. For leading zeros the joint probability that an animal is never captured before capture occasion t, and that it is alive in the population at t, is:

$$\psi_t = \psi_{t-1}(1 - p_{t-1})\phi_{t-1} + p_{ent(t)},$$

where:

- $\psi_1 = p_{ent(1)}$,
- ϕ_t is the probability that an animal survives from time t to time $t + 1$,
- $p_{ent(t)}$ is the probability that an animal enters the population for the first time at time t, and
- p_t is the probability that an animal is captured on capture occasion t, given that it is alive at the time.

For trailing zeros the conditional probability that an animal is never seen after t given that it is in the population at t is:

$$\chi_t = 1 - \phi_t + \phi_t(1 - p_{t+1})\chi_{t+1},$$

where $\chi_k = 1$, and where k is the number of capture occasions in the study. The capture history probability for animal i, that is first caught on occasion f and last caught on occasion l is:

$$P(x_i; \theta) = \psi_f \chi_l \prod_{t=f}^{l} p_t^{I_{t \in C}} (1 - p_t)^{I_{t \notin C}} \phi_t, \tag{4.2}$$

where C is the set of all occasions when animal i was captured, and:

$$I_{t\in C} = \begin{cases} 1, & t \in C, \\ 0, & t \notin C. \end{cases}$$

The detection probability is determined from its complement, the probability that an animal is never caught, which is given by:

$$1 - p_\theta = p_{ent(1)}(1 - p_1)\chi_1 + p_{ent(2)}(1 - p_2)\chi_2 + ... + p_{ent(k)}(1 - p_k)\chi_k.$$

The expected population sizes over time are given by:

$$E(N_1) = N_s p_{ent(1)},$$
$$E(N_t) = E(N_{t-1})\phi_{t-1} + N_s p_{ent(t)}.$$

The POPAN-lambda model (Carroll et al., 2013) constrains $p_{ent(t)}$ so that the population follows a smooth curve given by the growth rate

$$\lambda = \frac{E(N_{t+1})}{E(N_t)}.$$

4.2 Data simulation

We wrote R code to simulate capture histories given values for the relevant parameters of a POPAN model. Here are the first few histories from a dataset that was simulated with this code:

```
##   t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
## 1  0  0  1  0  0  0  0  0  0   0
## 2  0  0  0  0  1  0  0  0  0   0
## 3  0  0  0  1  0  0  1  0  0   0
## 4  1  1  0  0  0  0  0  0  0   0
## 5  0  0  0  1  0  0  0  1  0   1
## 6  0  1  1  0  0  0  0  0  0   1
```

Here are the life histories of the corresponding animals:

```
##   t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
## 1  1  1  1  1  1  1  1  1  1   1
## 2  1  1  1  1  1  0  0  0  0   0
## 3  1  1  1  1  1  1  1  1  1   1
```

```
## 4  1  1  0  0  0  0  0  0  0   0
## 5  1  1  1  1  1  1  1  1  1   1
## 6  1  1  1  1  1  1  1  1  1   1
```

Animals are listed in the order in which they enter the population, so the first capture histories are for animals that "entered" the population in the first survey. We specify a larger value for $p_{ent(1)}$ than for the other p_{ent} parameters because this corresponds to the first attempt to capture animals from an existing population, whereas later p_{ent} parameters reflect new births only.

Here are the final capture histories in the dataset:

```
##      t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
## 4383  0  0  0  0  0  0  0  0  0   1
## 4384  0  0  0  0  0  0  0  0  0   1
## 4385  0  0  0  0  0  0  0  0  0   1
## 4386  0  0  0  0  0  0  0  0  0   1
## 4387  0  0  0  0  0  0  0  0  0   1
## 4388  0  0  0  0  0  0  0  0  0   1
```

The final rows correspond to animals that entered the population at the last survey, and only capture histories for animals that were seen at least once are included. Here are the underlying life histories for these animals:

```
##      t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
## 4383  0  0  0  0  0  0  0  0  0   1
## 4384  0  0  0  0  0  0  0  0  0   1
## 4385  0  0  0  0  0  0  0  0  0   1
## 4386  0  0  0  0  0  0  0  0  0   1
## 4387  0  0  0  0  0  0  0  0  0   1
## 4388  0  0  0  0  0  0  0  0  0   1
```

We can check that animals are never captured when they are not in the population. Either they are alive or they are not captured, or both. Here true.pop = 1 if the animal is alive, and pop.dat = 1 if it is captured:

```
all(!pop.dat | true.pop)
```

```
## [1] TRUE
```

We can check that the population sizes over time are near the expected values:

```
##              t1   t2   t3   t4   t5   t6   t7   t8   t9  t10
## observed.N 2999 3444 3830 4173 4608 4972 5295 5474 5743 5936
## expected.N 3000 3478 3908 4295 4643 4957 5239 5493 5721 5927
```

We can check that the numbers captured at each survey are near the expected values.

```
##             t1  t2  t3  t4  t5  t6  t7  t8  t9 t10
## expected.C 480 413 345 793 737 547 424 985 402 890
## observed.C 485 424 348 802 742 558 412 980 413 899
```

Finally we can check that the total number of animals detected is near the expected value:

```
## expected.n observed.n
##       4347       4388
```

Our code draws the numbers of animals entering the population at each capture occasion from a multinomial distribution over the entry proportions and the superpopulation. It then loops over the capture occasions using fast, vectorised functions to enter the predetermined numbers of new animals into the population, and find which animals survive from the previous occasion, and which animals are captured from among those that are currently alive.

Our code is about five times faster than a previous version (CAPOW, Fewster et al., 2015), taking 2.2 seconds to simulate a small data set 5000 times, where the previous version takes 11.4 seconds. The average number of animals detected for both versions of the code is within 0.1% of the expected number.

One reason the new code is so much faster is that the old code draws the entry occasion for each animal individually in a loop, instead of drawing the numbers entering on each occasion all at once. It also draws a binomial capture outcome for all animals in the superpopulation at each capture occasion and discards those for animals that are not currently alive, instead of just drawing for those which are alive.

4.3 Model fitting

We wrote R code to fit POPAN models to capture history datasets and estimate the underlying parameter values, and tested the code using simulated populations. Here is an example of the estimates produced by our new code and the previously developed version:

```
##                      N   phi    p1    p2    p3    p4   p5 pent1 pent2
## estimates.new 1035.059 0.693 0.053 0.146 0.096 0.107 0.09 0.545 0.106
## estimates.old 1035.059 0.693 0.053 0.146 0.096 0.107 0.09 0.545 0.106
```

```
## true.values   1000.000 0.800 0.060 0.120 0.090 0.090 0.06 0.600 0.100
##               pent3 pent4 pent5
## estimates.new 0.188 0.044 0.117
## estimates.old 0.188 0.044 0.117
## true.values   0.100 0.100 0.100
```

All of the parameter estimates are equal up to 3 decimal places. Note that ϕ is assumed to be constant over time to help with parameter identifiability, which is a common problem for POPAN models (Fewster et al, 2015).

The log-likelihood for the POPAN model is:

$$l(N_s, \theta) = \log \binom{N_s}{n} + (N_s - n) \log(1 - p_\theta) + \sum_{i=1}^{n} \log P(x_i; \theta),$$

from (4.1), where:

$$\log P(x_i; \theta) = \log \psi_f + \log \chi_l + \sum_{t=f}^{l} \left(I_{t \in C} \log p_t + I_{t \notin C} \log(1 - p_t) + \log \phi_t \right)$$

from (4.2). It is not a regular likelihood because the observation n is truncated by the parameter N_s, and the binomial term means it is not a function only of i.i.d. observations. We fit the model using numerical optimisation.

Most of the work of evaluating the likelihood is in the capture history probabilities. Our code loops over the capture histories once, outside the negative log-likelihood function, storing the indices of the first and last captures and the subsequences of the capture histories between them. It takes frequency tables of the first and last captures and uses them to add the right number of the corresponding $\log \psi_t$ and $\log \chi_t$ terms in the negative log-likelihood function. It also takes column sums of the captures, and of non-captures between the first and last captures, and uses them to add the right numbers of the corresponding $\log p_t$, $\log(1 - p_t)$, and ϕ_t terms in the negative log-likelihood function. These are fast, vectorised functions, and their results can be added to the likelihood all at once.

The new model-fitting code is about four times faster than the previous version, taking 1.3 seconds to fit 50 small data sets where the previous version takes 5.2 seconds. The main reason the new code is faster is that the previous version finds the non-captures between the first and last captures inside the negative log-likelihood function. This means that they are found again every time the numerical optimiser evaluates the negative log-likelihood. The previous version also finds the probability of each capture history and adds it to the likelihood individually instead of using summaries to add all multiples of the same components at once. It also calculates ψ_t by recursing from the base case again for each t, instead of finding them all once.

We plan to use our new code for future releases and journal publication of the application mentioned in this section, a capture-recapture power-analysis package called CAPOW.

In the next section we describe how to simulate the additional genetic data required for close-kin methods.

Chapter 5

Realistic individual-level close-kin genetics data simulation for New Zealand southern right whales

Close-kin genetics is based on family relationships elicited from samples of genotyped animals. We are interested in applying the method to the New Zealand southern right whale (NZSRW) population (*Eubalaena australis*), which underwent a dramatic decline in abundance after an estimated 35,000 whales were slaughtered during 19th century whaling, and is now recovering. In this chapter we present R code for individual-level simulation under realistic scenarios for population and catch trajectories, life history traits, and genetic inheritance. All of these influence the distribution of family relationships, genetic diversity, and genotypes in a population. We believe that such detailed simulations are as yet uncommon, and will be useful in themselves as tools to investigate population genetics, as well as to produce data for developing new genetic modelling techniques. We also present a Shiny interface for viewing and validating the resulting datasets.

5.1 Population and catch trajectories

We produced population and catch trajectories using a previously-published analysis of "integrated population-level assessment of the whaling impact and pre-exploitation abundance" of the NZSRW (Jackson et al., 2016). This paper produces trajectories for NZSRW by exploring different assumptions about historical catch data and the assignment of catches to NZ or Australian stocks, and by integrating different data from the contemporary NZ population to constrain trajectories. In one scenario only whales caught in New Zealand waters are assumed to be from the New Zealand population, and in another scenario American ship-based whaling in Eastern Australian waters is also assumed to deplete the New Zealand population ("Southwest Pacific" catch history). High and low cases for catch from New

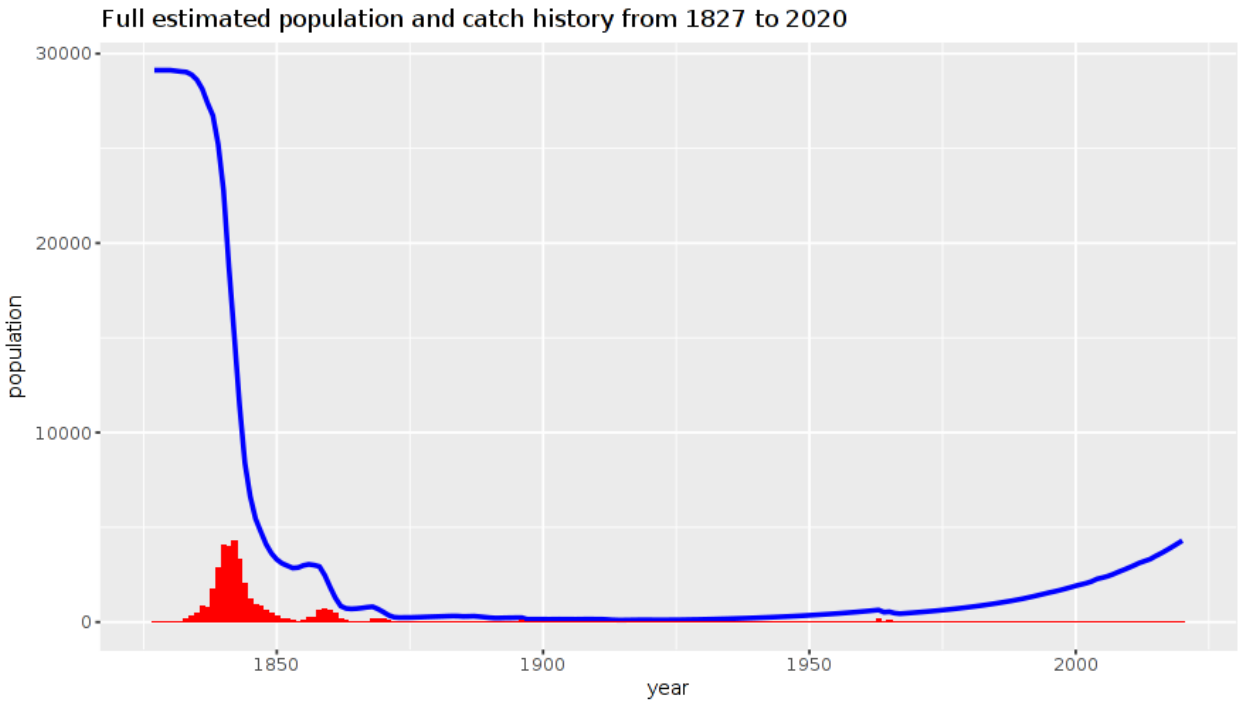


Figure 5.1: New Zealand scenario. Initial abundance estimated at approximately 30,000 whales.

Zealand shore-based and bay whaling are also distinguished. All of the catch history data are from Carroll et al. (2014).

Historical trajectories are constrained by using different bounds on the minimum historical population size, derived from the number of distinct mtDNA haplotypes observed in samples from the contemporary population (Jackson et al., 2008). Different trajectories are also produced using female capture-recapture data directly, and using abundance estimates from a POPAN model fitted to both male and female captures (Carroll et al., 2013).

We reproduced trajectories for the two scenarios described in the paper corresponding to the minimum and maximum total catch estimated. The first scenario was for New Zealand catches only, low case shore and bay catch, with a minimum population size of 36, using female capture data. The second scenario included Southwest Pacific catches and assumed high case shore and bay catch, with no minimum population size, using POPAN abundance estimates. We followed the R code posted with the paper, originally from Zerbini et al. (2011), generating samples from the posterior population and catch trajectories and taking their medians over time. In Figures 5.1 and 5.2 the solid blue lines represent the population, and the red histograms represent the catch. The scales of the y-axes are different between the two Figures, with the New Zealand scenario representing more conservative estimates of historical abundance. (All of the Figures in this section were taken from the shiny interface presented in subsection 5.4)

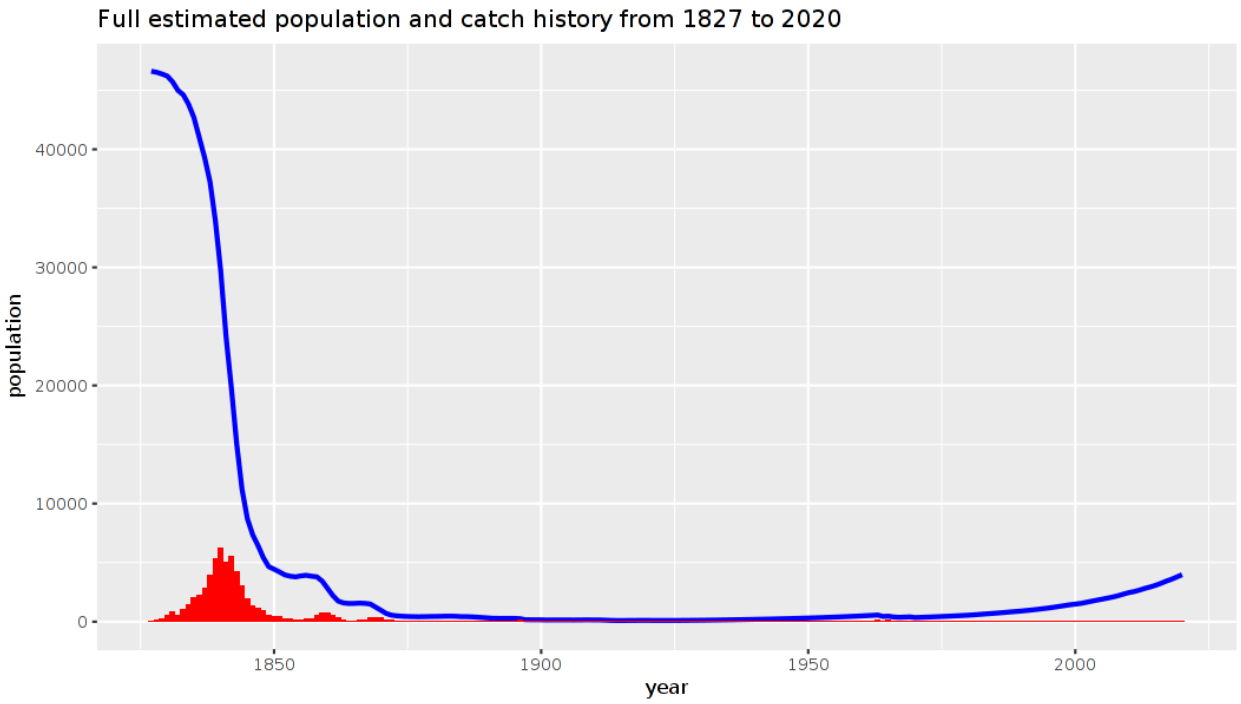


Figure 5.2: Southwest Pacific scenario. Initial abundance estimated at approximately 47,000 whales.

The previous POPAN data simulation we presented in Section 4.2 finds the numbers of animals entering the population (e.g. births) from the superpopulation parameter N_s and the entry proportions $p_{ent(t)}$. In contrast, here we use the population size at each time, N_t, the number caught c_t, and the natural survival probability assuming that the whale is not caught, ϕ, to create a population of individuals and their kinship relations based on the trajectories in Jackson et al. (2016). We assume that $\phi = 0.97$, constant over time and age based on empirical work in right whale populations and in common with Jackson et al. (2016). The latter is probably unrealistic for all demographic classes and we may extend the simulation to incorporate lower survival rates for calves in the future.

At the first year in the simulation (1827) we enter N_1 animals into the population. We then loop over the remaining years t, removing c_t animals selected with equal probability, finding those that survive death by other causes with probability ϕ, and entering enough new animals to satisfy N_t subject to certain constraints. If N_t turns out to be smaller than the number of animals surviving whaling and death by other causes then no new animals are entered and the difference is selected randomly and removed. We call this "emigration", and it often occurs for a few animals in the trajectories specified above during the period of most intensive whaling. Otherwise the constraints on the numbers of animals entering the population are due to knowledge of right whale breeding characteristics, as described below.

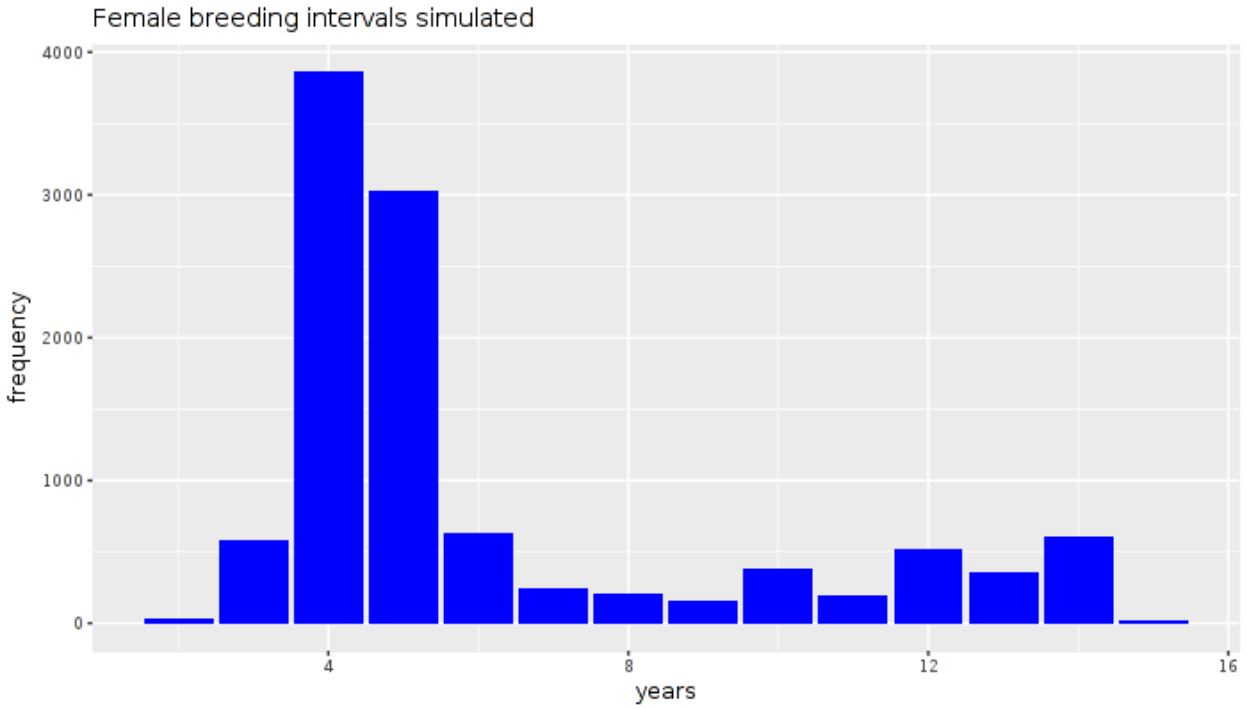


Figure 5.3: New Zealand scenario, longer breeding intervals on average.

5.2 Life history traits

To simulate data for close-kin genetics each new animal that is entered into the population must be assigned parents from which to inherit its genotype and family relationships. Doing this in a realistic way requires knowledge of the breeding characteristics of the species being simulated.

The age of sexual maturity for the SRW is suggested by the age at first parturition, which has been variously observed to have a mean of between 8.6 and 9.5 years (Charlton, 2017). We incorporate this into our simulation by specifying an age of maturity of $\alpha = 8$ years. We may extend our code to incorporate variability in age of maturity in the future.

Female SRWs calve at multi-year intervals due to the large investment of energy in gestation and nursing calves, which involves losing a significant proportion of their body weight (Christiansen et al 2018). We incorporate this into our simulation by specifying that females with the longest gap since last breeding are the first to breed in the current year. This causes breeding intervals for females to cluster around an average determined by the population and catch trajectories specified. This can be seen in the histograms below (Figures 5.3 and 5.4).

The long tails correspond to periods of reduced birthrate early in the trajectories when the population was near its equilibrium level. When ϕ is held constant and there is no whaling, minimal population growth implies low birthrates. For comparison, the histogram below

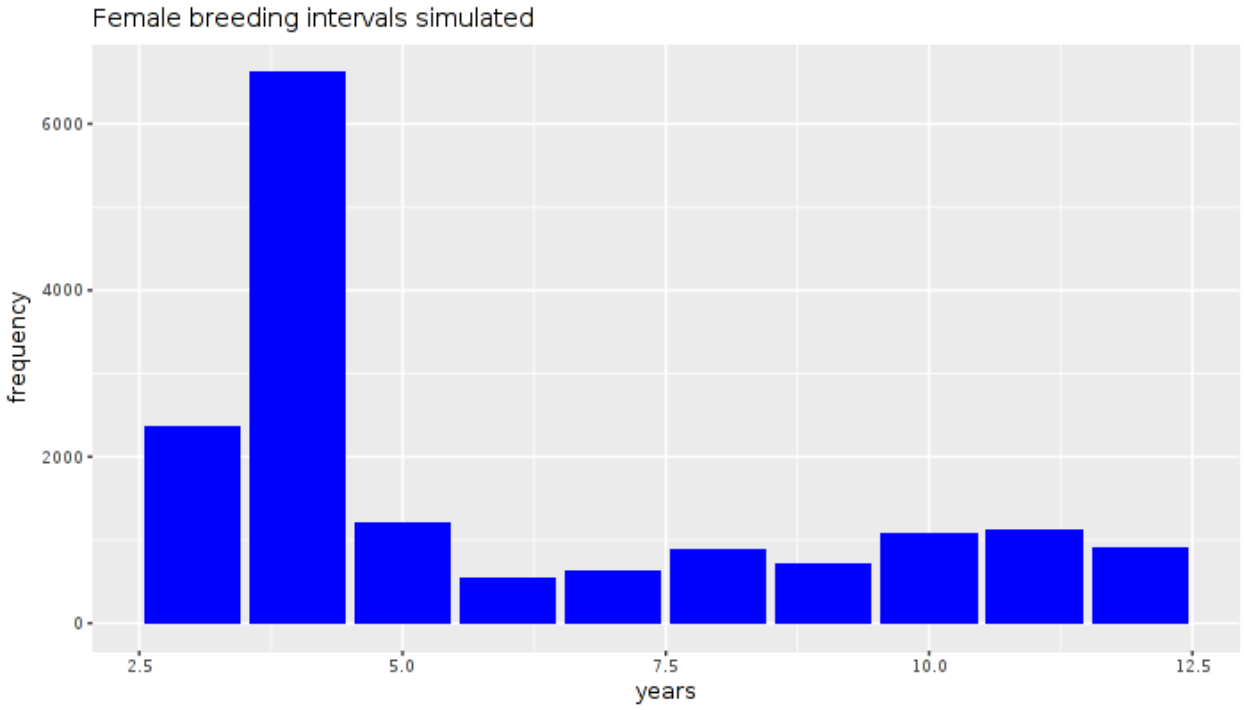


Figure 5.4: Southwest Pacific scenario, shorter breeding intervals on average.

(Figure 5.5) does not include that period.

The observed average breeding interval for southern right whale females is 3 years (IWC 2001), although observations specific to the New Zealand population are sparse, so the simulated intervals seem reasonable. The discrepancy could be due to atypically low population growth rates implied by the historical population and catch trajectories, which do not match those in current observation scenarios. Another possibility is that these low population growth rates were not due to long breeding intervals, but arose because a disproportionately high number of mature females were caught due to whaling occurring at calving grounds. Fewer mature females could imply a lower birthrate at the population level without longer individual breeding intervals. Currently the simulation does not incorporate variable catch probabilities among different demographic classes. Yet another possibility is that the age of maturity $\alpha = 8$ is set too low, implying that our simulated populations contain too many mature females, each breeding at longer intervals. It would be interesting to investigate these questions further using this simulation.

There is currently no strong evidence for selectivity in male breeding among SRWs and our simulation reflects this by selecting males for breeding with equal probability once they reach maturity. Thus male breeding intervals range over the natural numbers with decreasing frequency as interval-length increases, as can be seen in Figure 5.6.

After determining the number of animals required to enter the population we check whether there were enough mature animals in the previous year to be their parents. As baleen whales

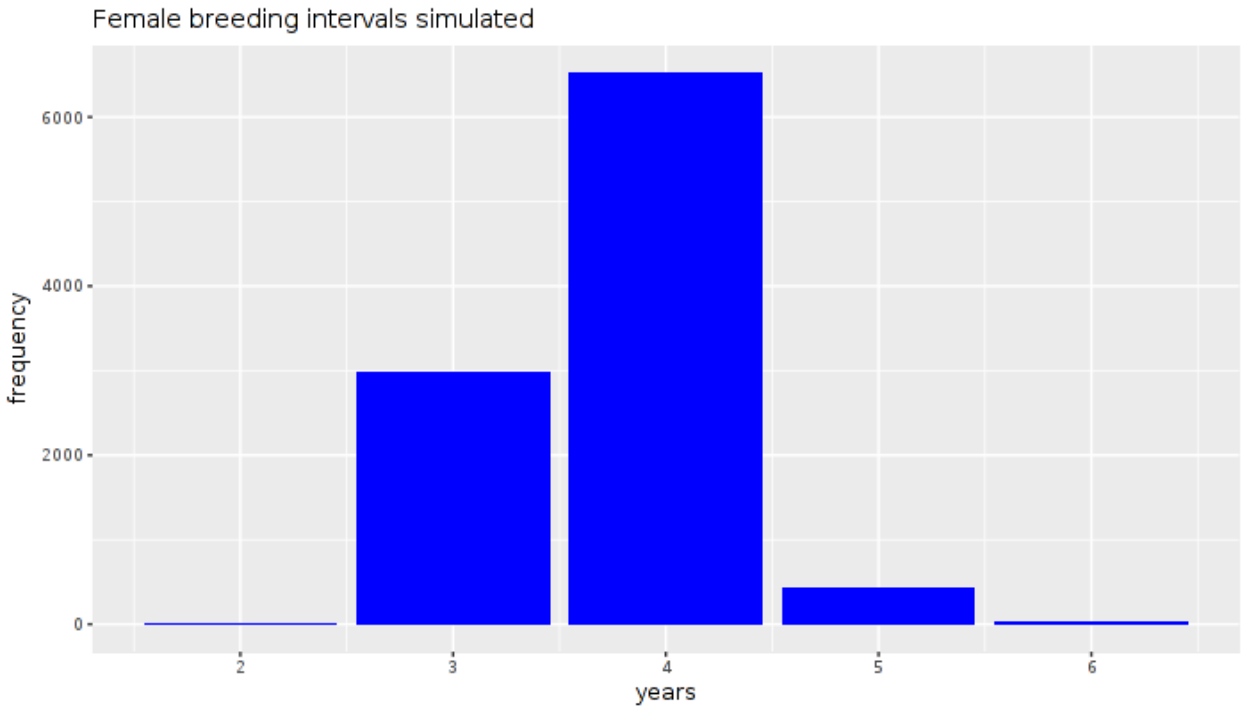


Figure 5.5: Southwest Pacific scenario from 1847 onwards after the major population crash.

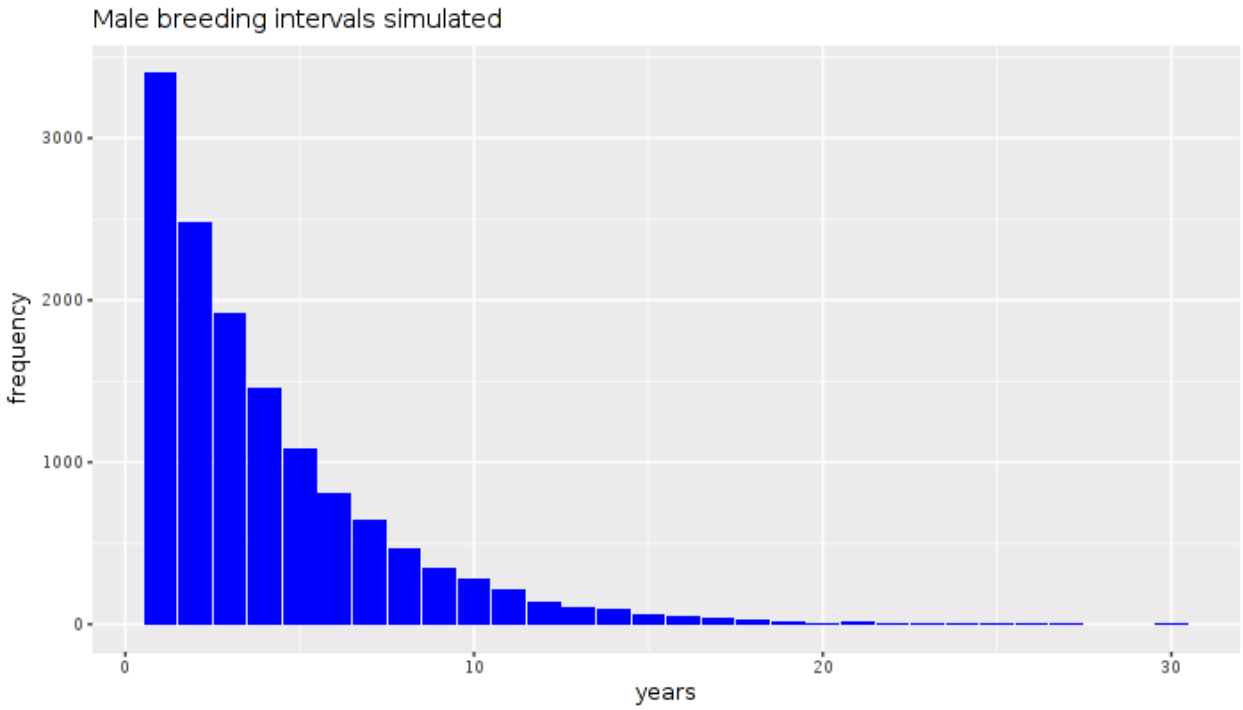


Figure 5.6: New Zealand scenario, male breeding intervals range over natural numbers.

such as SRWs very rarely give birth to twins, we require sufficiently many mature females to assign a unique mother to each new animal. There should also be at least one mature male, since males are not limited to fathering any maximum number of calves per year. If there are no mature males then no new animals are added. If there are fewer mature females than required we add one new animal for each mature female. If the number of animals added is consequently less than the required number, we record the loss and reduce the population sizes N_t to be simulated in the remaining years. The trajectories above are often reduced by a few animals in this way.

5.3 Genetic inheritance

We include three forms of genetic data in our simulation code: microsatellite (msat) DNA, mitochondrial DNA (mtDNA) haplotypes, and single nucleotide polymorphisms (SNPs). Diploid animals such as humans and whales have two sets of chromosomes where one allele at each locus is inherited maternally and one paternally. Msats and SNPs are inherited in this way. In contrast mtDNA is only inherited maternally as one locus of linked genes called a haplotype. There are many different haplotypes in a typical population, and several different possible alleles at a typical msat locus. In contrast there are usually only two, but theoretically up to four, different possible SNPs at one locus. Although this makes individual SNPs less informative, it is mitigated by the fact that it is easy to sequence large numbers of SNPs. We are interested in finding SNPs that are inherited together so that they can be considered as one more-informative locus with multiple possible alleles, so that is how we implement them here.

We implement each of these genotypes by initialising animals from a realistic distribution of observed msat and SNP alleles, or mtDNA haplotypes, based on previously published data from NZSRWs. For msats we use observed allele frequencies from 13 loci with between 6 and 21 alleles each, from Carroll et al. (2012). We use a multinomial distribution over the observed frequencies to generate genotypes for the initial population. Alleles are distinguished by their lengths. We can thus check that the initial population is allotted alleles at close to the observed relative frequencies by comparing the average length of alleles at each locus between simulated and observed data. Below are frequencies for the first few "maternally inherited" alleles, with suffix 'M', and the last few "paternally inherited" alleles, with suffix 'P' (although of course the initial population in the simulation does not inherit alleles at all.)

```
##   expectation.proportion    EV1_M   EV14_M   EV37_M GATA28_M
## 1            initialised 132.2167 134.3917 197.7167 171.1667
## 2              specified 131.5496 133.5701 198.1298 172.4015
```

```
##     RW18_P   RW31_P  RW410_P   RW48_P  TR3F4_P  TR3G1_P  TR3G2_P
## 1 196.5333 123.5167 202.6917 119.1833 313.7667 221.3667 175.6333
## 2 196.6533 123.4078 202.7989 119.0210 312.6996 222.9843 175.7974
```

We can also check that offspring inherit one allele randomly from each of their parents' genotypes at each locus. To demonstrate this Mendelian inheritance in the code, below is a randomly chosen whale, shown in rows labelled 1, and its mother in rows labelled 2, and father in rows labelled 3, and both alleles at each of the first few loci of their msat genotypes.

```
##     ID birth.year mum  dad female
## 1 2249       2000 810 1135      1
## 2  810       1987  NA   NA      1
## 3 1135       1991  NA   NA      0

##   EV1_M EV14_M EV37_M GATA28_M GATA98_M GT23_M RW18_M
## 1   138    131    203      174      116    116    195
## 2   138    131    203      174      116    106    187
## 3   122    141    195      174      108    116    215

##   EV1_P EV14_P EV37_P GATA28_P GATA98_P GT23_P RW18_P
## 1   126    137    195      178      116    116    215
## 2   158    141    203      174      104    116    195
## 3   126    137    189      178      116    110    195
```

The first few rows show the ID, birth year, parents' IDs, and gender of all three animals. We can see that the ID of the first animal is larger than those of the other two, and that it was born later than either of them. We can also see that the parents' IDs for the first animal are the IDs of the second and third animals. Note that these animals are from a simulation started after 1991 so that the parents were from the initial population and thus themselves had unknown parents. We can see that the first animal is female, and that its mother is female and its father is male.

The next few rows show the "maternally inherited" alleles at each of the first few loci of their msat genotypes, the column-names have the suffix 'M' to show this. The next few rows show the "paternally inherited" alleles at the same loci, the column-names have the suffix 'P' to show this. The maternally inherited alleles of the daughter have come from the mother, so each maternally inherited allele in the row labelled 1 occurs in the same locus as either the maternally or paternally inherited allele in one of the rows labelled 2. Likewise each paternally inherited allele in the row labelled 1 occurs in the same locus as either the maternally or paternally inherited allele in one of the rows labelled 3.

By comparing the daughter's maternal alleles with her mother's maternal and paternal alleles we can see that she inherited all of her mother's first 5 maternal alleles, and then two of her mother's paternal alleles. She inherited her father's maternal alleles at loci EV37, GT23, and RW18, and his paternal alleles at loci EV1, EV14, GATA28, and GATA98.

Haplotypes and SNPs are initialised in a somewhat more sophisticated way. We use allele frequencies observed for SRWs globally, together with established population genetics assumptions about selective breeding and migration, to generate realistic distributions for the

NZSRW. The mtDNA data are from Carroll et al. (2019), the SNP data are unpublished data from E L Carroll and O E Gaggiotti, and the code to produce the distributions was written by Emma Carroll in 2019. As these data are not used in the remainder of this dissertation we do not describe this process in detail here.

For simulating haplotypes, again we use a multinomial distribution over the resulting allele frequencies. We may check that the initial population contains each haplotype at close to the frequencies specified. Frequencies for the first and last few haplotypes are:

```
##   expectation.proportion  BakHapAA   BakHapBP   BakHapCC
## 1             initialised 0.1083333 0.03333333 0.05000000
## 2               specified 0.1510616 0.03236623 0.03802204

##     ValHappQ   ValHappR   ValHappW   ValHappX   ValHappY
## 1 0.00000000 0.02500000 0.00000000 0.00000000 0.00000000
## 2 0.00140483 0.00814618 0.00048165 0.00086199 0.00323898
```

We can also check that offspring inherit their haplotypes from their mothers. For the parent-offspring set above, we obtain:

```
##     ID birth.year mum  dad female haplotype
## 1 2249       2000 810 1135      1  PatHap17
## 2  810       1987  NA   NA      1  PatHap17
## 3 1135       1991  NA   NA      0  SA950028
```

As the SNPs in our simulations have exactly two alleles at each locus, each locus is initialised independently using a binomial distribution with its relative frequency from the distribution found as above. We can check the initial SNP distributions in the same way as for haplotypes. As for msats, there are maternal and paternal alleles for each animal:

```
##   expectation.proportion    SNP1_M     SNP2_M     SNP3_M
## 1             initialised 0.7750000 0.05833333 0.12500000
## 2               specified 0.7847453 0.05403802 0.09965725

##       SNP4_P    SNP5_P    SNP6_P     SNP7_P  SNP8_P    SNP9_P
## 1 0.05833333 0.1666667 0.4666667 0.07500000 0.35000 0.1166667
## 2 0.08195722 0.2274175 0.4454386 0.07034979 0.29998 0.1535949
```

All SNPs are assumed to be inherited altogether, like a mini-haplotype, but one from each parent, and with a large number of possible combinations. We can check that each set of maternally and paternally inherited alleles is a complete set of such alleles from the corresponding parent. Each SNP locus has two allele types denoted by 0 and 1, and there are ten loci altogether. Here are the first few of each for the same parent-offspring set as above:

```
##     ID birth.year mum  dad female
## 1 2249       2000 810 1135      1
## 2  810       1987  NA   NA      1
## 3 1135       1991  NA   NA      0

##   SNP1_M SNP2_M SNP3_M SNP4_M SNP5_M SNP6_M SNP7_M
## 1      1      0      0      0      0      0      0
## 2      1      0      0      0      0      0      0
## 3      1      0      0      0      0      0      0

##   SNP1_P SNP2_P SNP3_P SNP4_P SNP5_P SNP6_P SNP7_P
## 1      1      0      0      0      0      0      0
## 2      1      0      0      0      0      0      0
## 3      0      0      0      0      0      0      0
```

Here the mother has "inherited" the same set of alleles from both of her parents (again her genotype was initialised, not inherited) so we cannot tell which one the daughter inherited, but we can see that the daughter has inherited her father's maternal set of alleles. This simulation incorporating mini-haplotypes will be useful for future work when selecting and assessing panels of SNPs for close-kin genetic analyses. However, in the rest of this dissertation, SNPs are treated as individual loci and are assumed to be inherited independently, i.e., they are not linked.

5.4 Shiny interface

We believe that the simulation framework created here is uncommon in the extent to which it enables us to establish realistic population trajectories with individual-level genetic inheritance over time. We have created a web-based Shiny interface that allows simulation from either of the two population and catch trajectories described above. The rest of this dissertation focuses on genetic modelling techniques using data from this simulation framework, so for now the interface has remained primarily a tool for viewing and validating simulation results. We plan to extend the interface in the future to allow user-input trajectories, breeding characteristics, and initial genetic distributions.

As the current version stores and returns validation data for every animal observed over the entire simulation it runs slowly for realistic long-term trajectories due to the large numbers of animals simulated at the beginnings. We therefore allow the user to select starting points within the trajectories, which is a simple way to allow much shorter and faster simulations to be run when desired. The results are then analysed automatically, and important features are presented in tabular and graphical form.

In the html version of this dissertation the interface can be used below. Otherwise it can be viewed at https://catchit.stat.auckland.ac.nz/apps/genpopsim/.

Realistic individual-level close-kin genetics data simulation for New Zealand southern right whales

Population history:

Female captures, NZ-only, low catch - Jackson et al 2016

Starting year:

1827 1952 2017

1827 1847 1867 1887 1907 1927 1947 1967 1987 2007 2017

Full history may take several minutes including computing outputs after simulation

Simulate

Simulation | Full History | Microsatellite Loci | Haplotypes | SNPs | Description

Chapter 6

Saddlepoint approximations of pseudo log likelihood ratio distributions for close kinships

Close-kin genetics (Bravington et al., 2016; Hillary et al., 2018) is based on family relationships elicited from samples of genotyped animals. This involves calculating a test statistic between each pair of genotypes, which we call the pseudo log likelihood ratio (PLOD), for comparing the hypotheses that the pair has a specified closer kinship versus a specified less-close kinship: for example, a PLOD might compare the hypothesis that the pair is a parent-offspring pair with the hypothesis that the pair is unrelated. A threshold value of the PLOD is established, beyond which the pair is considered to have the closer of the two relationships. These relationship decisions are then carried forward into a population dynamics model to draw inferences about population size and demographics based on the observed kin relationships in the data. The method of predetermining kinships between each pair of animals, and deleting any pairs over which there is doubt, introduces a degree of subjectivity into the analysis. We hope to develop a new method that is instead based on a pseudo-likelihood which incorporates multiple levels of kinship together with kinship uncertainty. Instead of deciding threshold values for PLODs, our approach is to approximate the entire distribution of PLODs for each particular true kinship level. We do this using the saddlepoint approximation.

6.1 Conditional genopair probabilities given kinships

Genotypes comprise observed alleles at one or more loci in the genome of an individual. Depending on the study species and type of genetic marker under consideration, the level of genetic diversity of loci can be high, meaning there are many possible alleles at a given locus. Here we will only consider biallelic genetic markers called SNPs. Advances in genomic

sequencing technologies mean it is increasingly easier to survey large numbers of SNPs simultaneously and these markers are relatively easy to score, analyse and model given their digital/binary nature (Carroll et al., 2018).

If we call the different possible alleles at a particular locus A and B, then we can describe the possible genotypes for each animal at that locus as AA, AB, and BB. If we know that A and B have relative frequencies p and q, and make the common assumption that the population is in Hardy-Weinberg equilibrium (Andrews, 2010), then we can infer that the probabilities of each genotype for this animal at this locus are:

$$P(AA) = p^2, P(AB) = 2pq, P(BB) = q^2.$$

The genotypes for a pair of animals are together called a genopair, and there are nine possible outcomes for the two animals in the pair. The probabilities of the possible genopairs at one locus have been described with matrices where the rows and columns correspond to the possible genotypes of each of the two animals. For the genotypes above this would be:

$$P(g_i, g_j) \in \begin{bmatrix} P(AA, AA) & P(AA, AB) & P(AA, BB) \\ P(AB, AA) & P(AB, AB) & P(AB, BB) \\ P(BB, AA) & P(BB, AB) & P(BB, BB) \end{bmatrix}.$$

Here, g_i and g_j are the genotypes of animals i and j at this locus.

Genopair probabilities depend on the family relationships between the animals. When they are closely related they are more likely to have alleles at a locus that are *identical by descent* (IBD). For particular kinships between the animals there are particular probabilities for each possible number of IBD alleles at each locus.

The simplest kinship might be that the animals are "unrelated", in other words that their only family relationships are so distant as to be negligible, in which case we assume that they have no alleles IBD. The probabilities that they have zero, one, or two alleles IBD can usefully be described by the vector:

$$\kappa_{UP} = (1, 0, 0).$$

Here, the subscript "UP" denotes an unrelated pair. The genopair probabilities can then be determined by taking the probability of one genotype, and multiplying it by the conditional probability of the other given that no alleles are IBD, which is just its unconditional probability. The genopair probabilities are then given by:

$$G_{UP} = \begin{bmatrix} (p^2)(p^2) & (p^2)(2pq) & (p^2)(q^2) \\ (2pq)(p^2) & (2pq)(2pq) & (2pq)(q^2) \\ (q^2)(p^2) & (q^2)(2pq) & (q^2)(q^2) \end{bmatrix}$$

$$= \begin{bmatrix} p^4 & 2p^3q & p^2q^2 \\ 2p^3q & 4p^2q^2 & 2pq^3 \\ p^2q^2 & 2pq^3 & q^4 \end{bmatrix}. \tag{6.1}$$

Then $P\Big(g_i, g_j \Big| UP(i,j)\Big) \in G_{UP}$, where $UP(i,j)$ is the event that i and j are an unrelated pair.

Another important kinship is the parent-offspring pair. Because the offspring receives one of each of their parents' two alleles at each locus, the probabilities that they have zero, one, or two alleles IBD are therefore:

$$\kappa_{PO} = (0, 1, 0).$$

Here, the subscript "PO" denotes a parent-offspring pair. The genopair probabilities can then be determined by taking the probability of one genotype, and multiplying it by the conditional probability of the other genotype given that one allele is IBD:

$$G_{PO} = \begin{bmatrix} (p^2)(p) & (p^2)(q) & (p^2)(0) \\ (2pq)(\frac{p}{2}) & (2pq)(\frac{q}{2} + \frac{p}{2}) & (2pq)(\frac{q}{2}) \\ (q^2)(0) & (q^2)(p) & (q^2)(q) \end{bmatrix}$$

$$= \begin{bmatrix} p^3 & p^2q & 0 \\ p^2q & pq(p+q) & pq^2 \\ 0 & pq^2 & q^3 \end{bmatrix}. \tag{6.2}$$

A third important kinship is the self kinship. If i and j are actually the same animal caught twice, then the probabilities that they have zero, one, or two alleles IBD are:

$$\kappa_{SP} = (0, 0, 1).$$

The genopair probabilities are again the probabilities of one genotype, multiplied by the conditional probabilities of the other given that both alleles are IBD:

$$G_{SP} = \begin{bmatrix} (p^2)(1) & (p^2)(0) & (p^2)(0) \\ (2pq)(0) & (2pq)(1) & (2pq)(0) \\ (q^2)(0) & (q^2)(0) & (q^2)(1) \end{bmatrix}$$

$$= \begin{bmatrix} p^2 & 0 & 0 \\ 0 & 2pq & 0 \\ 0 & 0 & q^2 \end{bmatrix}.$$

These three kinships form a basis for the conditional genopair probabilities for all other kinships. An important example is a half-sibling pair. If two animals have one parent

in common then each of them gets one of that parent's two alleles at each locus. The probabilities that they share zero, one, or two alleles IBD at each locus are then:

$$\kappa_{HSP} = \Big(\frac{1}{2}, \frac{1}{2}, 0\Big).$$

The genopair probabilities can then be expressed as the probabilities of the numbers of alleles shared IBD multiplied by the conditional genopair probabilities given those IBD numbers at each locus, which are just the conditional probabilities for the three kinships above. For half-sibling pairs we have:

$$P\Big(g_i, g_j\Big|HSP(i,j)\Big) = \frac{1}{2}P\Big(g_i, g_j\Big|UP(i,j)\Big) + \frac{1}{2}P\Big(g_i, g_j\Big|PO(i,j)\Big). \tag{6.3}$$

We will describe formulations for other kinships in the next subsection.

6.2 Half-sibling vs unrelated pair pseudo log likelihood ratios

The test statistic which is the basis of close-kin genetics is the pseudo log likelihood ratio (called the PLOD in previous literature) for the hypotheses that the pair has a particular closer kinship versus a particular less-close kinship. For example we can take the closer kinship to be that the pair are half-siblings, and the less-close kinship to be that the pair are unrelated. Then the half-sibling versus unrelated pair PLOD is defined as:

$$PLOD^{HSP}_{UP}(i,j) = \frac{1}{n_L}\log\prod_{l=1}^{n_L}\frac{P\Big(g_i^l, g_j^l\Big|HSP(i,j)\Big)}{P\Big(g_i^l, g_j^l\Big|UP(i,j)\Big)}, \tag{6.4}$$

where n_L is the number of loci in the genotype, and g_i^l, g_j^l are the genotypes of individuals i and j at locus l.

If the loci were independent then the product over loci would equal the overall multi-locus genotype probabilities on both numerator and denominator. Denoting multi-locus genotypes by g_i and g_j, we would then have:

$$PLOD^{HSP}_{UP}(i,j) = \frac{1}{n_L}\log\frac{P\Big(g_i, g_j\Big|HSP(i,j)\Big)}{P\Big(g_i, g_j\Big|UP(i,j)\Big)}.$$

This is the log of the ratio of the likelihoods given by the observed genotypes to the hypotheses that i and j are half-siblings and unrelated respectively. It is an important result in statistical theory that a log likelihood ratio is the basis of the most powerful hypothesis

test at a given significance level. In close-kin genetics PLOD thresholds are used in this way to decide pairs of genotyped animals that have certain family relationships. However usually loci are not all truly independent, as some of them may be linked physically on the same chromosome, which is why the expression above is called a pseudo log likelihood ratio. Researchers make assessments which loci in a study are truly independent.

The relative magnitudes of $\prod_{l=1}^{n_L} P\Big(g_i^l, g_j^l \Big| HSP(i,j)\Big)$ and $\prod_{l=1}^{n_L} P\Big(g_i^l, g_j^l \Big| UP(i,j)\Big)$ imply the sign and magnitude of $PLOD_{UP}^{HSP}(i,j)$, which approximates the direction and strength of evidence from the genopair that the animals are half-siblings versus unrelated. PLODs greater than zero suggest more evidence for the animals being half-siblings.

We also have:

$$PLOD_{UP}^{HSP}(i,j) = \frac{1}{n_L}\sum_{l=1}^{n_L} \log \frac{P\Big(g_i^l, g_j^l \Big| HSP(i,j)\Big)}{P\Big(g_i^l, g_j^l \Big| UP(i,j)\Big)},$$

and

$$\frac{P\Big(g_i^l, g_j^l \Big| HSP(i,j)\Big)}{P\Big(g_i^l, g_j^l \Big| UP(i,j)\Big)} = \frac{1}{2}\left\{1 + \frac{P\Big(g_i^l, g_j^l \Big| PO(i,j)\Big)}{P\Big(g_i^l, g_j^l \Big| UP(i,j)\Big)}\right\},$$

from (6.3),

$$\in \frac{1}{2}\left\{1 + \begin{bmatrix} p^{-1} & (2p)^{-1} & 0 \\ (2p)^{-1} & (p+q)(4pq)^{-1} & (2q)^{-1} \\ 0 & (2q)^{-1} & q^{-1} \end{bmatrix}^l\right\},$$

from (6.2) and (6.1), where the superscript l on the matrix implies the genopair probability matrix for the l-th locus. This shows that the PLOD is larger when the animals share rare alleles, and smaller when they share no alleles, intuitively implying evidence that they are more or less closely related respectively.

With enough genetic data these PLODs form a discrete but fine-grained distribution. We used the simulation code described in Section 5 to simulate data for a population of New Zealand southern right whales from 1827 to 2000, with the NZ-scenario population and catch trajectories. We simulated genotypes consisting of 1000 independent, biallelic SNP loci, and randomly sampled 5% of the population per year from 1991 to 2000. The calculations for loci with more than two possible alleles are quite different and it would be interesting to compare them but we do not do so in this dissertation. Figure 6.1 shows the distribution of half-sibling versus unrelated pair PLODs among the sample of genotyped animals thus simulated.

We can see that the overwhelming majority of PLODs are less than zero, implying evidence that the corresponding pairs of genotyped animals are unrelated rather than half-siblings.

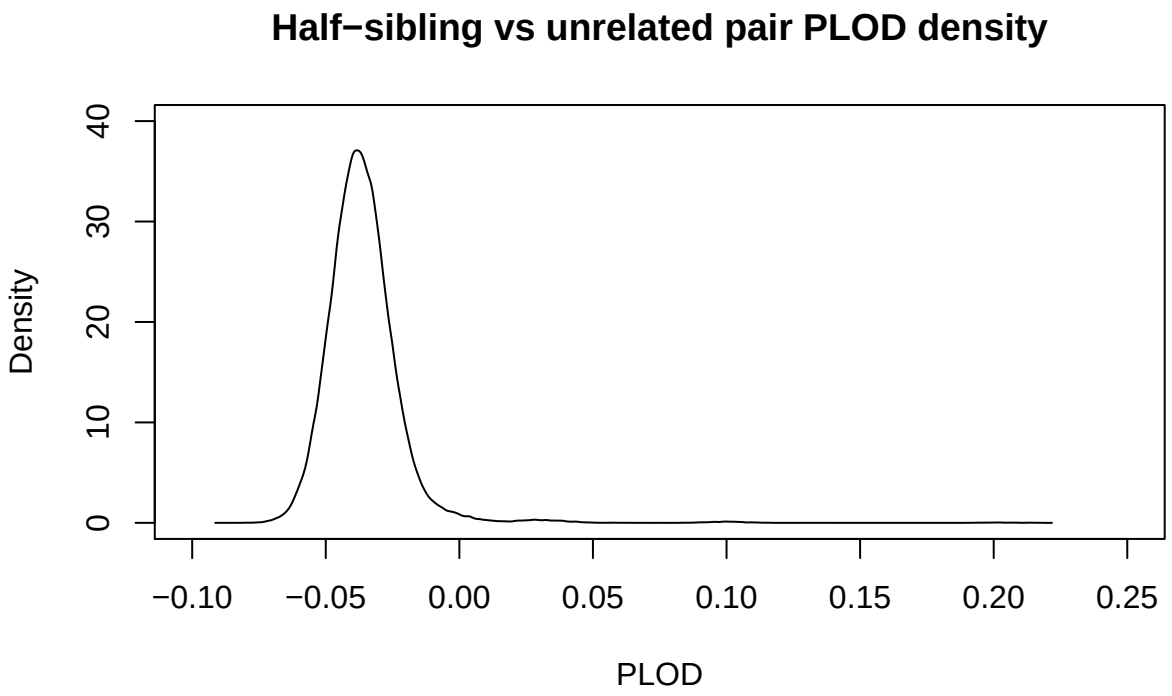


Figure 6.1: Full distribution of 277885 PLODs (all pairs).

Greater detail can be seen by zooming in on the base of the same curve as in Figure 6.2. There we see three distinct bumps centered at approximately 0.03, 0.1, and 0.2. As these bumps correspond to PLODs that are greater than zero, the corresponding pairs of genotypes are more likely to be obtained if the animals in the pair are half-siblings than if they are unrelated. In fact they correspond to sets of PLODs for pairs of animals that are (1) grandparent-grandchild or half-siblings, (2) parent-offspring, and (3) self-pairs respectively, as we show in the next subsection.

6.3 PLOD distributions given kinships

We are interested in describing the distribution of PLODs shown in Figures 6.1 and 6.2. If we can compute the theoretical probability density of the PLOD for a given set of population parameters, corresponding to the true underlying probability density curve that is approximated by the kernel density estimates in these figures, then we can use it to describe the pseudo-likelihood of the observed PLOD values in the data. Because they are pairwise comparisons, the observed PLODs are not independent and identically distributed (eg. the PLOD of a pair (a, b) affects the probability function of a PLOD (b, c)) so only a pseudo-likelihood can be computed. This proposed pseudo-likelihood incorporates multiple levels of kinship together with kinship uncertainty. Our approach is to partition the population of genopairs by kinship, approximate the probability density of the PLOD for each kinship separately, and then combine those approximations to compile the overall PLOD probability

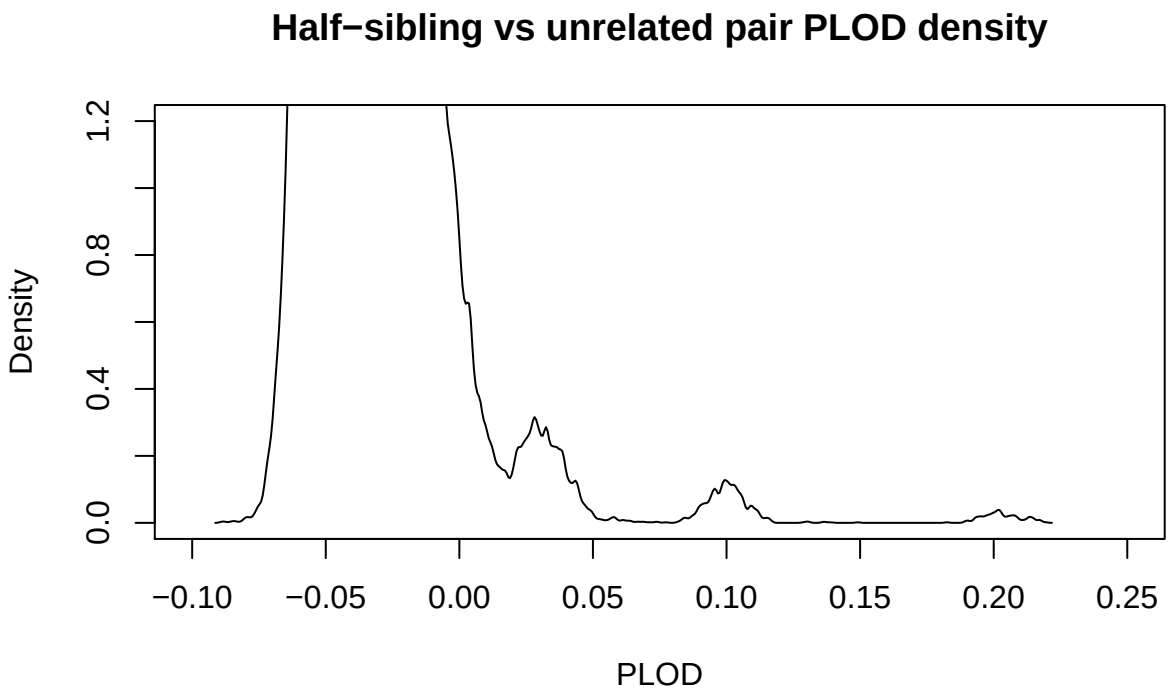


Figure 6.2: Base of distribution of 277885 PLODs (all pairs).

density function.

We use tidyverse database functions in R to efficiently find kinship pairs for animals in our simulated sample based on the parents of each animal recorded during the simulation. Self-pairs are simply the number of recaptures for each animal. Parent-offspring pairs are found by joining duplicated data tables on the ID column of one and the parent column of the other. Half-siblings require joining tables by mother and father and selecting pairs with one ID greater than the other to exclude duplicate pairs and self-pairs. Grandparents for each animal are found by selecting the parents of the parents. Then grandparent-grandchild pairs are found by joining tables on ID and grandparents. Thiatic (aunt/uncle to niece/nephew) pairs are found by joining tables on parents and grandparents. First cousin pairs are found by joining on grandparents but excluding siblings and self-pairs.

Mixed kinships are possible but uncommon. Kinships such as full-siblings are not common among SRWs because they are not monogamous. The kinships described above are thus enough to develop our methods.

Figure 6.3 shows the empirical distribution of PLODs from simulated data, with kernel density curves and sample means plotted separately for pairs satisfying each of the kinships listed above. The specific PLOD used in Figure 6.3 corresponds to the log-likelihood ratio for hypotheses of half-sibling and unrelated pairs, but we see that this PLOD also distinguishes multiple additional kinship-groups.

Intuitively, the less closely related the animals are, the less they will appear from their genotypes to favour the half-sibling hypothesis versus the unrelated hypothesis, and the

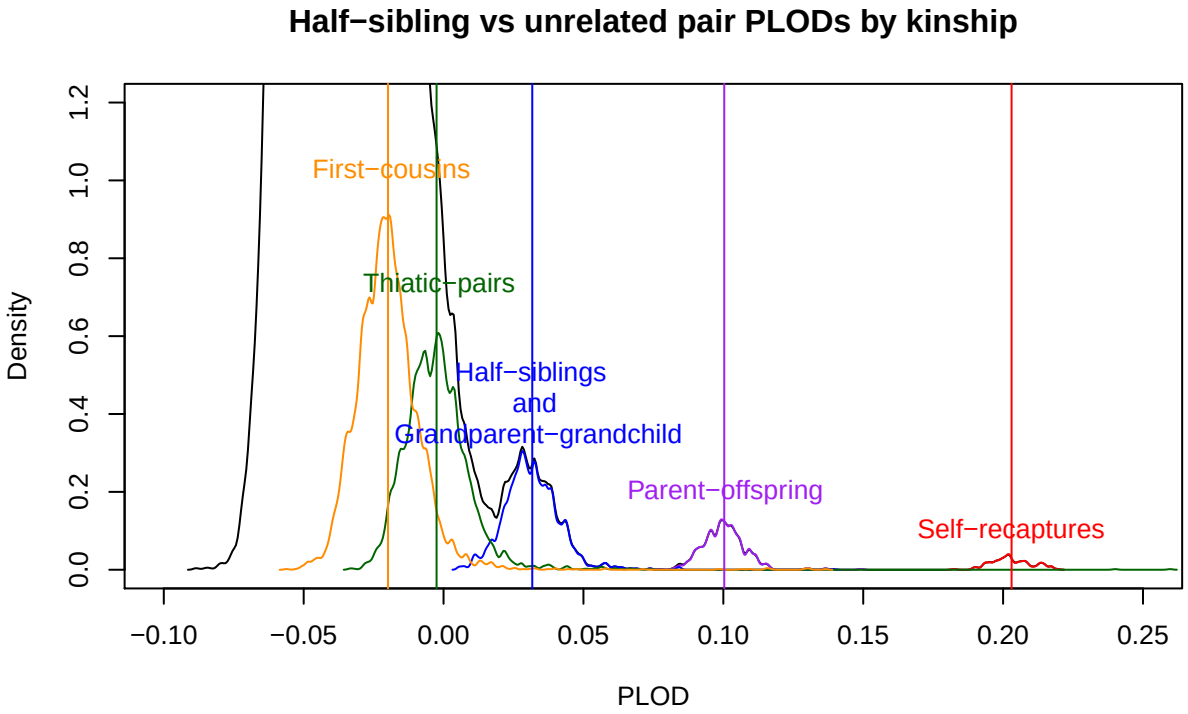


Figure 6.3: Distributions and sample means of PLODs separated by kinship.

smaller the associated PLOD will be. We now describe this systematically in terms of the relationship between the conditional genopair probabilities given the kinships of the animals, and the number of parent-offspring links separating them.

The conditional probabilities of genopairs given more remote kinships can be described in terms of those for unrelated and parent-offspring pairs as described in Section 6.1, although the formulation becomes a little more complicated. For illustration, take grandparent-grandchild pairs (grand-pairs), where one of the grandparent's offspring is one of the grandchild's parents. That offspring has one of the grandparent's two alleles at each locus, and the grandchild inherits that allele with probability 0.5 at each locus. The probabilities that they share zero, one, or two alleles IBD at each locus are then:

$$\kappa_{GP} = \Big(\frac{1}{2}, \frac{1}{2}, 0\Big),$$

where "GP" refers to grand-pair, and the genopair probabilities can be expressed as:

$$P\Big(g_i, g_j\Big|GP(i,j)\Big) = \frac{1}{2}P\Big(g_i, g_j\Big|UP(i,j)\Big) + \frac{1}{2}P\Big(g_i, g_j\Big|PO(i,j)\Big),$$

exactly as for half-sibling pairs in (6.3). This is why grand-pair PLODs are plotted together with half-sibling PLODs on Figure 6.3 above.

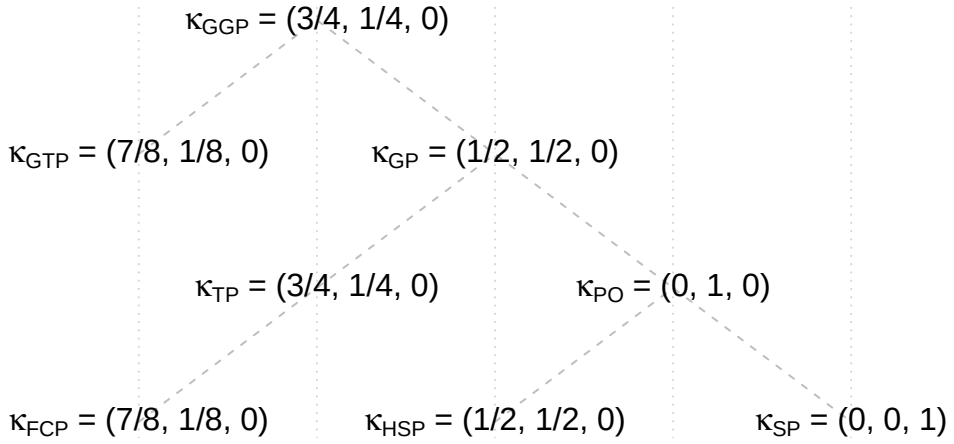


Figure 6.4: A kind of family tree where each diagonal line represents a parent-offspring link, and each node shows the genopair probabilities for the pair composed of that family member and the one at the bottom right, expressed as κ vectors as explained in Section 6.1.

For genotypes in which one allele is inherited from each parent at each locus, each parent-offspring link in a kinship after the first reduces the probability that the animals have one allele IBD by a factor of two, and the amount lost from the probability of one allele IBD is added to the probability that they have no alleles IBD at each locus. We created a diagrammatic visualization of this process that makes it easy to see how to formulate the κ vector for different relationships: see Figure 6.4.

The kinships referred to in Figure 6.4 are:

- SP: Self-pair
- PO: Parent-offspring pair
- HSP: Half-sibling pair
- GP: Grand-pair
- TP: Thiatic-pair
- GGP: Great grand-pair
- FCP: First-cousin pair
- GTP: Grand thiatic-pair

6.4 Saddlepoint approximations

It is difficult to determine the probability of any particular value of a PLOD because of the computationally intractable number of possible genopairs, 9^{n_L} since there are 9 genopairs at each locus, and the number of loci n_L must be sufficiently large for the genotypes to be informative. Our approach is to use the saddlepoint approximation (Butler, 2007), which is a technique for generating accurate approximations to a probability function given the

associated moment generating function. A probability mass function for a random variable X can be approximated by the saddlepoint density:

$$\tilde{f}(x) = \frac{1}{\sqrt{2\pi|K''(\hat{s})|}} \exp\{K(\hat{s}) - \hat{s}x\}, \tag{6.5}$$

where

$$K(s) = \log[E\{\exp(sX)\}],$$

the log of the moment generating function of X, and where $\hat{s}$ satisfies the saddlepoint equation $K'(s) = x$.

The moment-generating function for the half-sibling versus unrelated pair PLOD given a certain kinship KS is:

$$E[\exp\{sPLOD_{UP}^{HSP}(i,j)\}|KS(i,j)] = E\Big(\exp\Big[s\frac{1}{n_L}\sum_{l=1}^{n_L}\log\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}\Big]\Big|KS(i,j)\Big),$$

from (6.4),

$$= E\Big(\prod_{l=1}^{n_L}\exp\Big[s\frac{1}{n_L}\log\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}\Big]\Big|KS(i,j)\Big)$$

$$= \prod_{l=1}^{n_L} E[\exp\{sPLOD_{UP}^{HSP}(i,j)^l\}|KS(i,j)],$$

where the superscript implies the contribution to the overall PLOD from the l-th locus, assuming independent loci.

Letting $P_{ij} = PLOD_{UP}^{HSP}(i,j)$ we then have

$$K(s) = \sum_{l=1}^{n_L}\log E\{\exp(sP_{ij}^l)|KS(i,j)\},$$

$$K'(s) = \sum_{l=1}^{n_L}\frac{E\{P_{ij}^l\exp(sP_{ij}^l)|KS(i,j)\}}{E\{\exp(sP_{ij}^l)|KS(i,j)\}},$$

and

$$K''(s) = \sum_{l=1}^{n_L}\Big\{\frac{E\{(P_{ij}^l)^2\exp(sP_{ij}^l)|KS(i,j)\}}{E\{\exp(sP_{ij}^l)|KS(i,j)\}} - \Big[\frac{E\{P_{ij}^l\exp(sP_{ij}^l)|KS(i,j)\}}{E\{\exp(sP_{ij}^l)|KS(i,j)\}}\Big]^2\Big\}.$$

All of the expectations here are weighted averages over the nine conditional genopair probabilities given kinship KS, at locus l. For example:

$$E\{\exp(sP_{ij}^l)|KS(i,j)\} =$$

$$\sum_{g_i^l}\sum_{g_j^l} P\{g_i^l, g_j^l|KS(i,j)\}\exp\Big[s\frac{1}{n_L}\log\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}\Big]$$

$$= \sum_{g_i^l}\sum_{g_j^l} P\{g_i^l, g_j^l|KS(i,j)\}\Big[\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}\Big]^{\frac{s}{n_L}}.$$

The saddlepoint, $\hat{s}$, is found by numerical minimisation of $K(\hat{s}) - \hat{s}x$, providing the exponent in the expression for the approximated density (6.5) in the process.

The vertical lines on Figure 6.5 correspond to the expected value of the PLOD itself for pairs of animals with kinship KS, which is given by:

$$E\{P_{ij}|KS(i,j)\} = E\Big[\frac{1}{n_L}\sum_{l=1}^{n_L}\log\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}\Big]$$

$$= \frac{1}{n_L}\sum_{l=1}^{n_L} E\Big[\log\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}\Big]$$

$$= \frac{1}{n_L}\sum_{l=1}^{n_L}\sum_{g_i^l}\sum_{g_j^l} P\{g_i^l, g_j^l|KS(i,j)\}\log\frac{P\{g_i^l, g_j^l|HSP(i,j)\}}{P\{g_i^l, g_j^l|UP(i,j)\}}.$$

We wrote R code to calculate the saddlepoint densities and expected values of the PLODs for all of the kinships above. We used the sample allele frequencies to approximate the genotype allele probabilities for each animal, and checked that this gave results very close to using the frequencies among the entire breeding population in the simulation when the animals were conceived. Results are shown in Figure 6.5, with the approximated densities for each kinship gained from equation (6.5), and the partition weights for each kinship gained from the observed numbers of corresponding pairs in the sample. The smooth saddlepoint curves exhibit a good fit to the kernel densities of the sample PLOD values for each kinship category.

Figure 6.6 gives the combined density for all of the kinships isolated above, together with the corresponding combined saddlepoint approximation.

Using saddlepoint approximations for the distributions of PLODs for sufficiently many major kinships, we can approximate the entire distribution of PLODs with an average weighted by the probabilities of each kind of kinship pair. In the next section we express several such kinpair probabilities in terms of our parameters of interest, and in Section 8 we describe the pseudo-likelihood thus created.

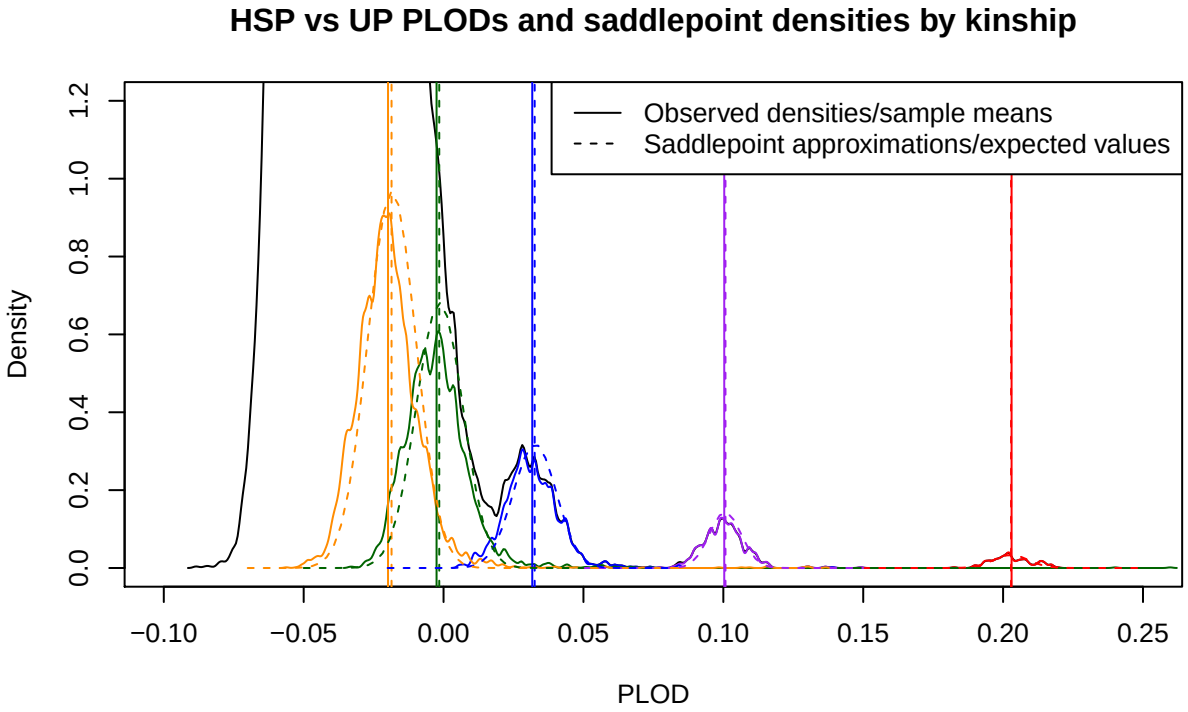


Figure 6.5: The smooth dashed lines are the saddlepoint approximations to the PLOD PDF for each kinship, and the solid jagged lines are the sample-based kernel densities.

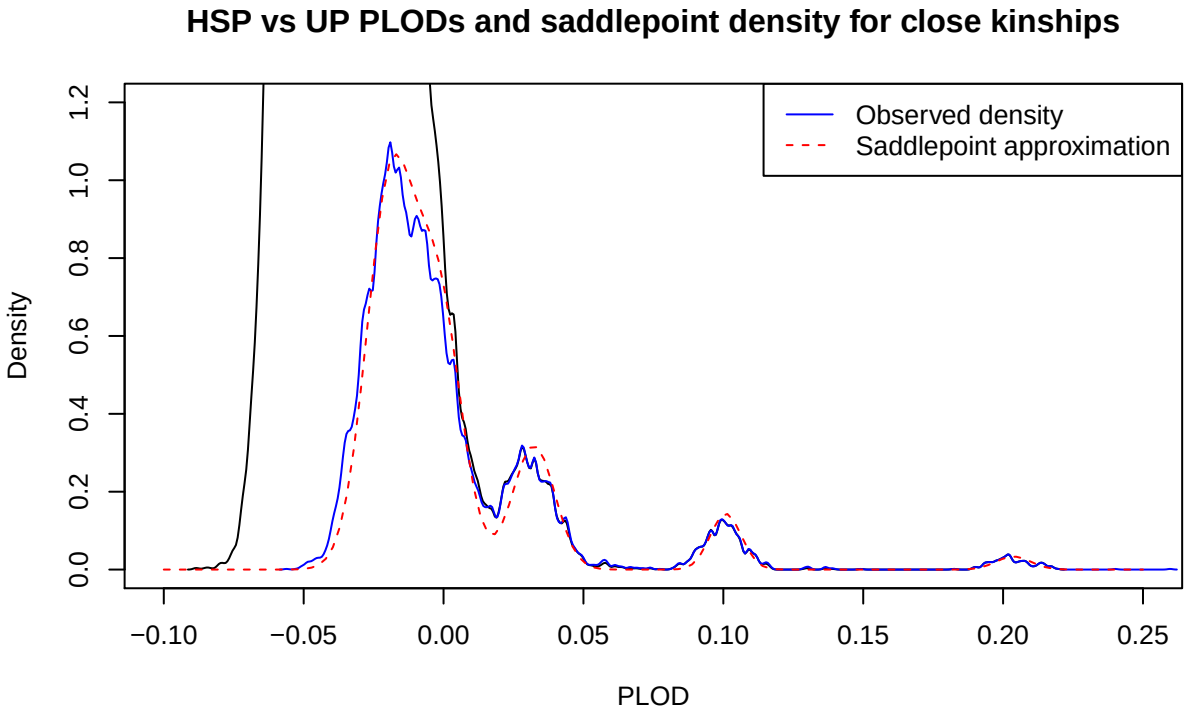


Figure 6.6: The smooth dashed line is the saddlepoint approximation to the PLOD PDF for all of the kinships above, and the solid jagged line is the sample-based kernel density.

Chapter 7

Self and parent-offspring pair probabilities within and between samples

Our new method is based on a pseudo-likelihood which incorporates multiple levels of kinship together with kinship uncertainty. Our approach to finding the probability density function of the PLOD is to partition over kinship level. We approximate the probability density function of PLODs within each kinship using the saddlepoint method, and combine the densities for different kinships by weighting them according to the probabilities of the corresponding kinships given our parameters of interest. This requires us to find the probability that a randomly-chosen pair of genotypes has a particular kinship level, from within the sample space of all pairs of genotyped samples. We call this probability the *kinpair* probability. We develop our method by deriving expressions for three kinpair probabilities: parent-offspring pairs among genotypes from one sampling occasion, self-pairs among genotypes from different sampling occasions, and parent-offspring pairs among genotypes from sampling occasions that are separated by a period which is less than the age of maturity. We show that the numbers of each of these kinpairs occurring in datasets generated using our simulation framework are distributed around those predicted by these expressions.

While developing our method we make several simplifying assumptions. We assume that the survival probability ϕ and the population growth rate λ are constant over time, that there is no excess mortality due to whaling, and no migration. We also assume that the birthrate, $\lambda - \phi$, is greater than zero, which implies that $\phi < \lambda$.

Our parameters are then $\theta = \{N_t, \lambda, \phi\}$, where N_t is the population size at time t, and:

$$\lambda = \frac{E(N_{t+1})}{E(N_t)},$$

as defined in Section 4.

7.1 Parent-offspring pairs within one sampling year

The relationship between the expected population sizes at years t_1 and t_2 is given by:

$$E(N_{t_1}) = \frac{E(N_{t_2})}{\lambda^{t_2-t_1}}. \tag{7.1}$$

The probability that an animal survives from year t_1 until year t_2 is:

$$P(S_{t_1,t_2}) = \phi^{t_2-t_1}. \tag{7.2}$$

The expected number of animals born in year t is the expected number that were alive the year before, multiplied by the birth rate:

$$E(B_t) = E(N_{t-1})(\lambda - \phi)$$

$$= E(N_t)\frac{(\lambda - \phi)}{\lambda}. \tag{7.3}$$

The expected number of parent-offspring pairs among animals alive in the population at a survey year t_s is the expected number of animals born each year up to and including t_s, multiplied by the probability that each one survives until year t_s, and by the probability for each of its parents that it survives from the year before the animal was born (when it was conceived) until year t_s. Writing $\sum_{t=t_s}^{-\infty}$ to denote summation over a decrementing index t, this gives:

$$E(PO_{t_s}) = \sum_{t=t_s}^{-\infty} E(B_t)P(S_{t,t_s})2P(S_{t-1,t_s})$$

$$= \sum_{t=t_s}^{-\infty} E(N_t)\frac{(\lambda - \phi)}{\lambda}\phi^{t_s-t}2\phi^{t_s-t+1},$$

from (7.2) and (7.3),

$$= 2\frac{\phi(\lambda - \phi)}{\lambda}\sum_{t=t_s}^{-\infty}\frac{E(N_{t_s})}{\lambda^{t_s-t}}\phi^{2(t_s-t)},$$

from (7.1),

$$= 2E(N_{t_s})\frac{\phi(\lambda - \phi)}{\lambda}\sum_{t=t_s}^{-\infty}\left(\frac{\phi^2}{\lambda}\right)^{t_s-t}$$

$$= 2E(N_{t_s})\frac{\phi(\lambda-\phi)}{\lambda}\left\{\left(\frac{\phi^2}{\lambda}\right)^0 + \left(\frac{\phi^2}{\lambda}\right)^1 + ...\right\}$$

$$= 2E(N_{t_s})\frac{\phi(\lambda-\phi)}{\lambda}\frac{1}{1-\frac{\phi^2}{\lambda}},$$

as $0 < \phi < 1$ and $\phi < \lambda$ implies that $0 < \frac{\phi^2}{\lambda} < 1$,

$$= 2E(N_{t_s})\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}. \tag{7.4}$$

The number of pairs of animals in a population of size $E(N_{t_s})$ is the number of combinations of size 2. We write this as $E(AP_{t_s})$:

$$E(AP_{t_s}) = \frac{E(N_{t_s})!}{\{E(N_{t_s})-2\}!2!},$$

for $E(N_{t_s}) \geq 2$,

$$= \frac{E(N_{t_s})\{E(N_{t_s})-1\}}{2}.$$

The probability that a pair of animals i and j that are alive in a population of size $E(N_{t_s})$ is a parent-offspring pair is given by:

$$P\{PO_{t_s}(i,j)|\theta\} = \frac{E(PO_{t_s})}{E(AP_{t_s})}$$

$$= \frac{4}{\Big(E(N_{t_s})-1\Big)}\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}.$$

This PO probability decreases with the expected size of the population, which is the key intuition behind close-kin genetics. The nature of the relationship between the PO probability and λ and ϕ is more complicated.

7.2 Self pairs between sampling years

If we consider pairs of animals in which one is alive in the population at t_1, and one at t_2, where $t_1 < t_2$, then the expected number of self-pairs available for sampling is just the expected number of animals that survive from t_1 until t_2:

$$E(SP_{t_1,t_2}) = E(N_{t_1})P(S_{t_1,t_2})$$

$$= E(N_{t_1})\phi^{t_2-t_1},$$

from (7.2).

The total number of pairs available for comparison from populations of size $E(N_{t_1})$ and $E(N_{t_2})$ respectively is:

$$E(AP_{t_1,t_2}) = E(N_{t_1})E(N_{t_2}). \tag{7.5}$$

We can then express the probability that a pair of animals i and j drawn from populations of size $E(N_{t_1})$ and $E(N_{t_2})$ respectively, is a self-pair, given our parameters, as:

$$P\{SP_{t_1,t_2}(i,j)|\theta\} = \frac{E(SP_{t_1,t_2})}{E(AP_{t_1,t_2})}$$

$$= \frac{\phi^{t_2-t_1}}{E(N_{t_2})},$$

$$= \frac{1}{E(N_{t_1})}\left(\frac{\phi}{\lambda}\right)^{t_2-t_1},$$

for $t_1 < t_2$, from (7.1). This self-pair probability increases with survival probability, and decreases with the population growth rate, the expected population size, and the length of the interval, as we would expect.

7.3 Parent-offspring pairs between samples separated by less than the age of maturity

Again considering pairs of animals in which one is alive at t_1 and one is alive at t_2, where $t_1 < t_2$, then the expected number of parent-offspring pairs can be partitioned into those that include animals that are born between t_1 and t_2, and those that do not. For those that do not, the parent and offspring must both be alive at time t_1, and each of these two animals that survives to time t_2 creates an additional parent-offspring pair between the samples; so the expected number of pairs between the samples is the expected number of parent-offspring pairs at t_1 multiplied by twice the survival probability to time t_2. For those that do, if the interval between t_1 and t_2 is not longer than the age of maturity α, then any animal born between t_1 and t_2 must have two parents that were both alive at t_1, and any parent-offspring pair from samples taken at times t_1 and t_2 must consist of offspring at time t_2 and parent

at time t_1, because the offspring was not alive at time t_1. The expected number of parent-offspring pairs between the samples is then twice the expected number that are born between t_1 and t_2 and which survive until t_2, constituting two parents for each of the offspring alive at time t_2. Putting the two expressions together gives:

$$E(PO_{t_1,t_2}) = E(PO_{t_1})2P(S_{t_1,t_2}) + 2\sum_{t=t_1+1}^{t_2} E(B_t)P(S_{t,t_2}),$$

for $t_2 - t_1 \leq \alpha$,

$$= 2E(N_{t_1})\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}2\phi^{t_2-t_1} + 2\sum_{t=t_1+1}^{t_2} E(N_t)\frac{(\lambda-\phi)}{\lambda}\phi^{t_2-t},$$

from (7.2), (7.3), and (7.4),

$$= 2\frac{E(N_{t_2})}{\lambda^{t_2-t_1}}\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}2\phi^{t_2-t_1} + 2\frac{(\lambda-\phi)}{\lambda}\sum_{t=t_1+1}^{t_2}\frac{E(N_{t_2})}{\lambda^{t_2-t}}\phi^{t_2-t},$$

from (7.1),

$$= 2E(N_{t_2})\Big\{2\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1} + \frac{(\lambda-\phi)}{\lambda}\sum_{t=t_1+1}^{t_2}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t}\Big\}$$

$$= 2E(N_{t_2})\Big[2\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1} + \frac{(\lambda-\phi)}{\lambda}\Big\{\Big(\frac{\phi}{\lambda}\Big)^{0} + ... + \Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1-1}\Big\}\Big],$$

$$= 2E(N_{t_2})\Big[2\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1} + \frac{(\lambda-\phi)}{\lambda}\left\{\frac{1-\left(\frac{\phi}{\lambda}\right)^{t_2-t_1}}{1-\left(\frac{\phi}{\lambda}\right)}\right\}\Big]$$

as $\phi < \lambda$,

$$= 2E(N_{t_2})\Big\{2\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1} + 1 - \Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1}\Big\}$$

$$= 2E(N_{t_2})\Big[\Big\{2\frac{\phi(\lambda-\phi)}{\lambda-\phi^2} - 1\Big\}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1} + 1\Big].$$

We can then express the probability that such a pair of animals i and j is a parent-offspring pair given our parameters, as the expected number of such parent-offspring pairs divided by the number of all such pairs in populations of sizes $E(N_{t_1})$ and $E(N_{t_2})$:

$$P\{PO_{t_1,t_2}(i,j)|\theta\} = \frac{E(PO_{t_1,t_2})}{E(AP_{t_1,t_2})}$$

$$= \frac{2}{E(N_{t_1})}\Big[\Big\{2\frac{\phi(\lambda-\phi)}{\lambda-\phi^2}-1\Big\}\Big(\frac{\phi}{\lambda}\Big)^{t_2-t_1}+1\Big],$$

for $0 < t_2 - t_1 \leq \alpha$, from (7.5). Again, this decreases with the expected population size, and has more complicated relationships with λ and ϕ.

7.4 Comparison with simulation

We used the simulation framework described in Section 5 to check the accuracy of the expressions derived above. We used a simplified population trajectory satisfying our assumptions above, with $\phi = 0.97$, $\lambda = 1.02$, and no catch trajectory or migration, for 165 years from 1827 to 1991, with a final population size of $N_{1991} = 2000$. We ran this full simulation ten times, and for each one we simulated ten sample histories, making 100 in total. In each sample history we randomly sampled one third of the population at each of six years, at five year intervals, from 1966 to 1991. We used the R code described in Section 6.3 to find the numbers of each of the three kinpairs described above. To simplify our code we only searched for self-pairs for samples separated by less than the age of maturity, namely for consecutive samples in the sample histories above, for which we could also check our expression for parent-offspring pairs. We compared the numbers found with those predicted by our expressions above. Figures 7.1 - 7.3 are histograms of the differences observed as proportions of the predicted numbers.

The observed numbers of kinship pairs from our simulations are approximately normally distributed around the expected numbers derived above, verifying our calculations. In the next section we combine the kinpair probabilities from this section with the saddlepoint approximations from the previous section, to form a pseudo-likelihood for observed PLODs.

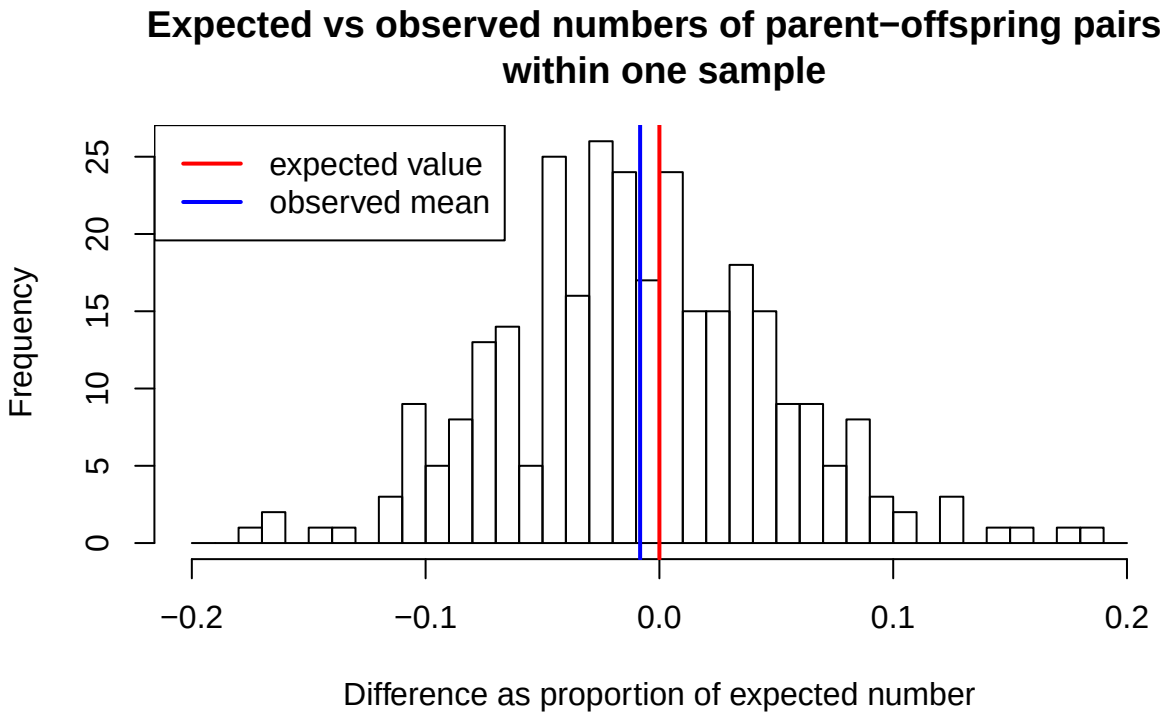


Figure 7.1: The average observed number of kinpairs is within 1% of the expected number.

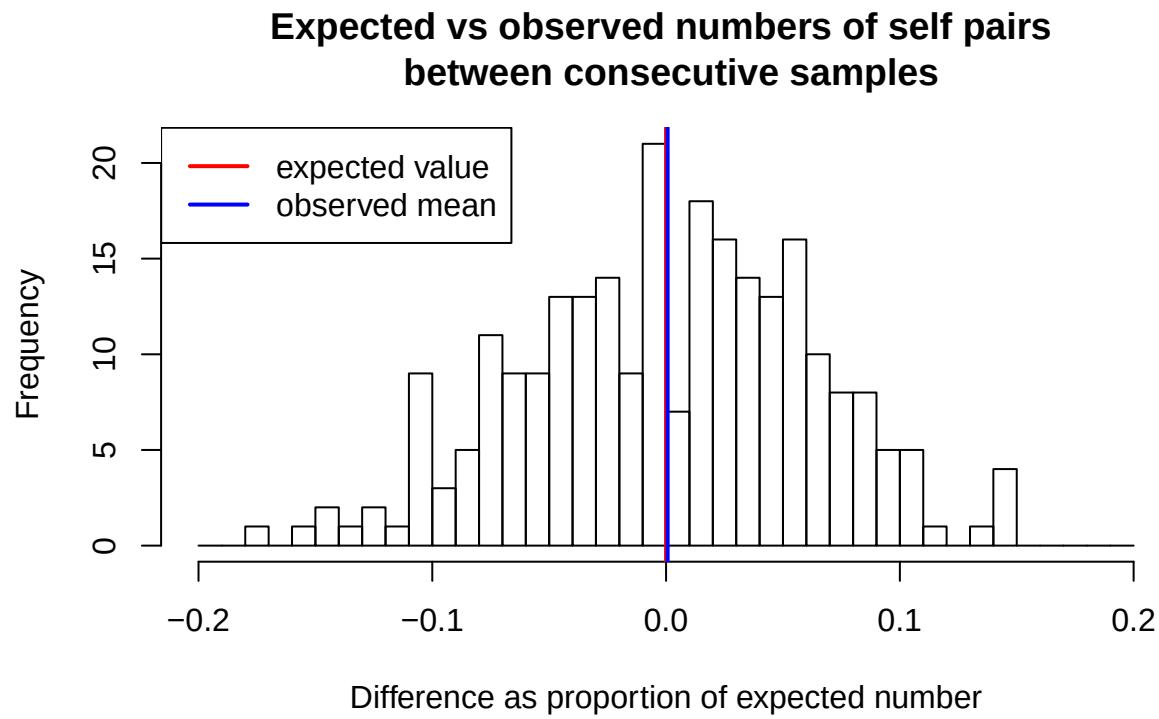


Figure 7.2: The average observed number of kinpairs is very close to the expected number.

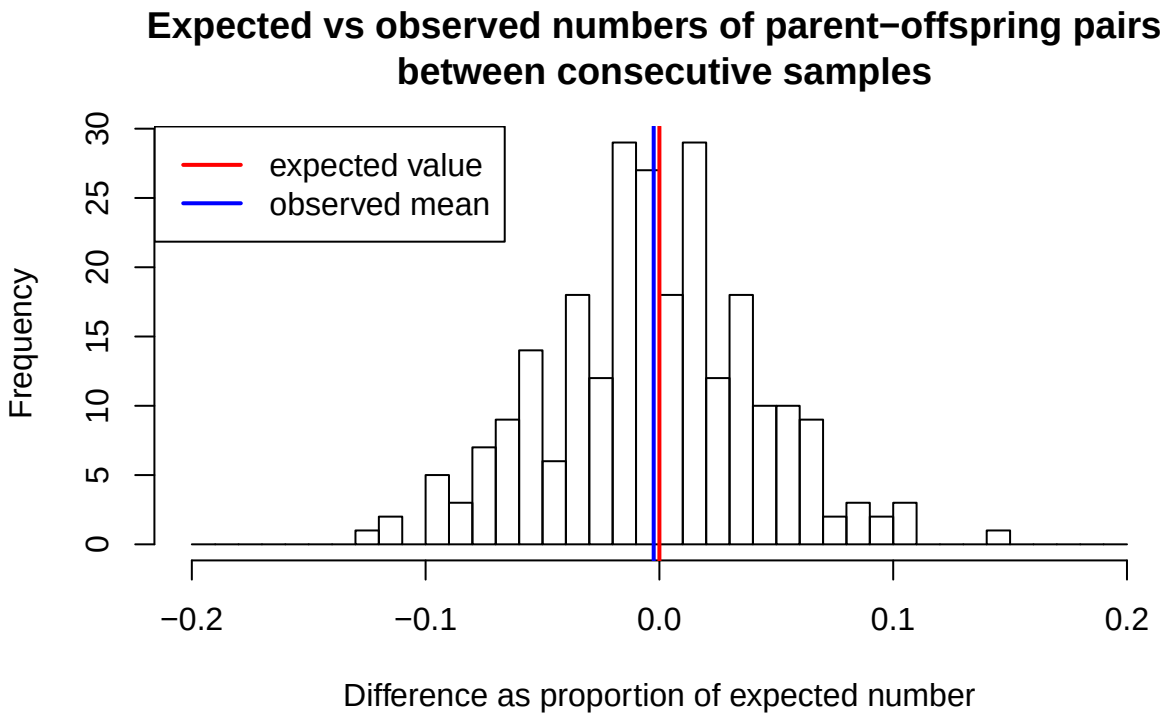


Figure 7.3: The average observed number of kinpairs is very close to the expected number.

Chapter 8

New method for estimating population size and demographics

Our new method forms a pseudo-likelihood for observed PLODs by combining saddlepoint approximations of PLOD distributions for pairs of animals with particular kinships, weighted by the corresponding kinpair probabilities given our parameters of interest. Here we present a preliminary analysis using just the three kinpair probabilities derived in section 7. We show that the resulting pseudo-likelihood can be maximised over our parameters of interest using numerical methods. We use our simulation framework to show that even this preliminary model can produce usefully precise estimates of abundance, survival and growth rate from realistic sample sizes and numbers of samples.

8.1 Pseudo-likelihood for parent-offspring versus unrelated pair PLODs

We have derived kinpair probabilities for parent-offspring and self-pairs so we use the parent-offspring versus unrelated pair PLOD, because it is more appropriate for separating the former pairs from less-related pairs. In the future we would like to compare this to an alternative of using the parent-offspring versus half-sibling pair PLOD, as half-siblings are the closest-related kinpairs after parent-offspring. We thus let $P_{ij} = PLOD^{PO}_{UP}(i,j)$.

For samples $s_1, ..., s_n$ of genotyped animals alive in the population at years t_1, ..., t_n, our proof-of-concept pseudo log-likelihood is given by:

$$l(s_1, ..., s_n; \theta) = \sum_{k=1}^{n} \sum_{i,j \in s_k, i \neq j} \log \Big[P\{PO_{t_k}(i,j)|\theta\} \tilde{f}\{P_{ij}|PO_{t_k}(i,j)\} + P\{\neg PO_{t_k}(i,j)|\theta\} \tilde{f}\{P_{ij}|\neg PO_{t_k}(i,j)\} \Big] + \quad (8.1)$$

where $\neg KS(i,j)$ is the event that animals i and j do not have kinship KS, and $\tilde{f}\{P_{ij}|KS(i,j)\}$ is the value for animals i and j of the saddlepoint approximation of the probability mass function of their PLOD given that they have kinship KS. We note that the probability of the value of a PLOD depends on the parameters θ only through the kinpair probabilities. That is, given kinship, the PLOD probabilities $P\{P_{ij}|KS(i,j)\}$, here approximated by $\tilde{f}\{P_{ij}|KS(i,j)\}$, are invariant to the parameters for population size, survival, and population growth rate.

The likelihood is thus an application of the law of total probability. For simplicity we do not include PLODs between samples separated by more than the age of maturity α, even though our expression for self-pair probabilities also applies to these.

We do not have expressions for $\tilde{f}\{P_{ij}|\neg PO_{t_k}(i,j)\}$ or $\tilde{f}\{P_{ij}|\neg PO_{t_k,t_l}(i,j) \cap \neg SP_{t_k,t_l}(i,j)\}$ because we do not have the underlying conditional genopair probabilities given those kinships. However, empirical work suggests that when using the parent-offspring versus unrelated pair PLOD for sufficiently informative genotypes, the distributions of PLODs for parent-offspring and less related pairs are quite distinct. The probabilities of PLODs for less related pairs, and their saddlepoint approximations, are very close to zero wherever the probabilities of PLODs for parent-offspring and self pairs are not very close to zero, and vice versa. That is, for every pair i,j, we find that for example either $\tilde{f}\{P_{ij}|PO_{t_k}(i,j)\} \approx 0$, or $\tilde{f}\{P_{ij}|\neg PO_{t_k}(i,j)\} \approx 0$. We can use this to show that the values of $\tilde{f}\{P_{ij}|\neg PO_{t_k}(i,j)\}$ do not affect the likelihood maximisation. Using the inner sum of the first term on the right-hand side of (8.1) for an example, the contribution to the pseudo log-likelihood is:

$$\begin{aligned}
&\sum_{i,j} \log\Big[P\{PO_{i,j}|\theta\}\tilde{f}\{P_{ij}|PO_{i,j}\} + P\{\neg PO_{i,j}|\theta\}\tilde{f}\{P_{ij}|\neg PO_{i,j}\}\Big] \\
&\quad\approx \sum_{i,j:\tilde{f}\{P_{ij}|\neg PO_{i,j}\}\approx 0} \log\Big[P\{PO_{i,j}|\theta\}\tilde{f}\{P_{ij}|PO_{i,j}\}\Big] + \\
&\qquad \sum_{i,j:\tilde{f}\{P_{ij}|PO_{i,j}\}\approx 0} \log\Big[P\{\neg PO_{i,j}|\theta\}\tilde{f}\{P_{ij}|\neg PO_{i,j}\}\Big] \\
&\quad= \sum_{i,j:\tilde{f}\{P_{ij}|\neg PO_{i,j}\}\approx 0} \log\Big[P\{PO_{i,j}|\theta\}\tilde{f}\{P_{ij}|PO_{i,j}\}\Big] + \\
&\qquad n^* \log P\{\neg PO_{i,j}|\theta\} + n^* \bar{E}\Big[\log \tilde{f}\{P_{ij}|\neg PO_{i,j}\}\Big],
\end{aligned}$$

where $i,j \in s_k, i \neq j$, $PO_{i,j} = PO_{t_k}(i,j)$, n^* is the number of pairs i,j such that $\tilde{f}\{P_{ij}|PO_{i,j}\} \approx 0$, and $\bar{E}\Big[\log \tilde{f}\{P_{ij}|\neg PO_{i,j}\}\Big]$ refers to the sample mean of the log approximated PLOD probabilities of those pairs.

The form of this expression, together with similar working for the other terms in the pseudo log-likelihood, shows that the values of the parameters at which the likelihood is maximised do not depend on the $\tilde{f}\{P_{ij}|KS_{i,j}\}$ terms when these densities are sufficiently well separated for different kinships. Since we do not have expressions for each of the kinship levels as yet, for the time being we choose a constant value to represent those that are missing, for example $\tilde{f}\{P_{ij}|\neg PO_{i,j}\}$ for those pairs i, j for which this is non-zero in the example above.

Indeed, the $\tilde{f}\{P_{ij}|KS_{i,j}\}$ terms for parent-offspring and self-pairs are not strictly necessary either in this situation: we are using them to illustrate our method. In fact, when all kinship levels are distinguishable it is possible to find the maximum likelihood estimates of the parameters analytically, without saddlepoint approximation, but our goal is to avoid the requirement that all kinships used in an analysis should be distinct from one another. Incorporating all kinships, without much more informative genotypes than are currently available, requires extending our method to include expressions for all of the kinpair probabilities in terms of our parameters, which we hope to do in the future.

We describe our approach as a pseudo-likelihood because the observations consist of all pairwise comparisons, and are therefore not independent even if the DNA samples are. For example, if an observed genotype has large PLODs with two other observed genotypes then it is much more likely that those two other observed genotypes have a large PLOD with each other. Also, as we mentioned in Section 6.4 we are assuming in our saddlepoint approximations that the loci are independent, when often they are not, due to linkage disequilibrium.

However, under certain conditions, for example if our sample is sufficiently small relative to the population size, if enough of our loci are independent, and if our saddlepoint approximations and allele probability approximations are sufficiently accurate, our pseudo-likelihood might approximate a genuine likelihood well. Indeed, these are the circumstances under which close kin genetics has been successfully applied in the past (e.g., tuna and shark work). In that case, unlike for POPAN and other capture-recapture models for estimating population size, our pseudo-likelihood would approximate a regular likelihood, consisting of approximately independent, identically distributed observations, not depending on sample size or other observations for which the support depends on the parameters of interest. This can be seen from the fact that the parameters of interest only weight the distributions of PLODs for different kinships; they do not change their supports. While this depends on a lot of non-trivial conditions, it is an exciting possibility because of the many desirable statistical properties of regular likelihoods.

8.2 Maximum likelihood estimates

We wrote R code to calculate the pseudo-likelihood described above and find the maximum likelihood estimates using numerical optimisation and grid search. We found that two of the most popular optimisation functions in R, optim and nlm, both produced inconsistent results. We thus used starting points in a grid surrounding the true values, as well as

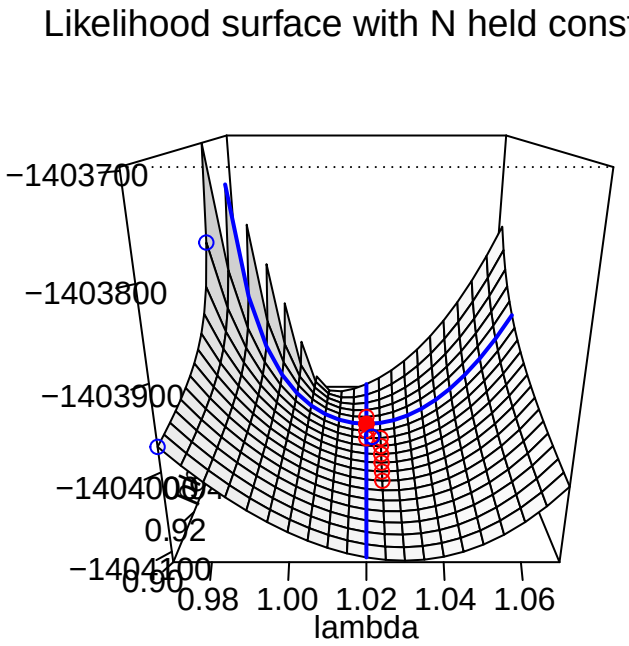


Figure 8.1: Grid search has found the true values of λ and ϕ. The results from two of the starting points for optim are close.

implementing grid search throughout, and took the maximum likelihood estimates overall. We would like to try starting the optimisers from the best estimates from grid search but we do not expect the resulting estimates to be much better.

We used the allele frequencies of the breeding population at the time of conception for each animal to estimate the allele probabilities. In practice we would only be able to use the sample allele frequencies, but for exploring the performance of our new ideas we aimed to eliminate any bias from our preliminary results that might arise from subsidiary processes such as genetic drift.

Figures 8.1 - 8.3 show the negative log pseudo-likelihood surfaces from grid searches with one parameter at a time held constant at its true value, for a sample history (a set of samples at particular times for a particular population) created using our simulation framework as described in Section 7.4. The blue lines show the true values of the parameters, the blue circles show the values returned by optim when started from the four corners of the parameter space, the hollow red circles show the ten best results from grid search, and the solid red circles show the best result from grid search in each case.

We can see that the negative log-likelihood surface is nonlinear but generally smooth and continuous. The exception to this is an asymptotic increase (signifying a decreasing likelihood) at the vertical plane $\lambda = \phi$ bounding the parameter space in the far left corner of Figure 8.1. This bound may be one reason that the numerical optimisers perform inconsistently, as it requires the use of a less stable numerical method ("L-BFGS-B" in the case of optim).

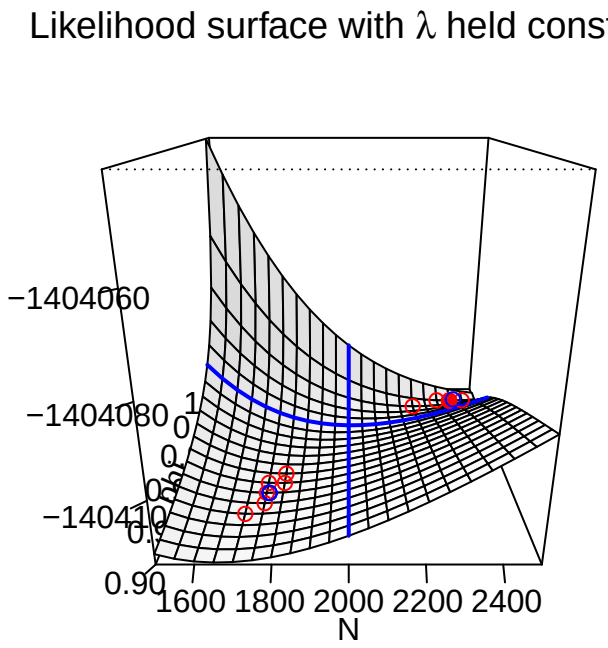


Figure 8.2: Both grid search and optim find that the MLEs are relatively far from true values of N and ϕ, and the likelihood is bi-modal.

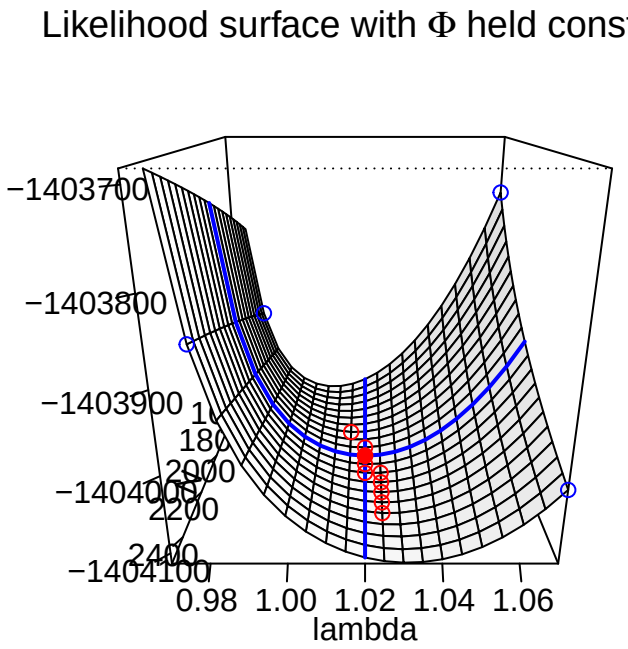


Figure 8.3: Grid search has found the true values of λ and N. None of the estimates from optim have moved far from their starting points.

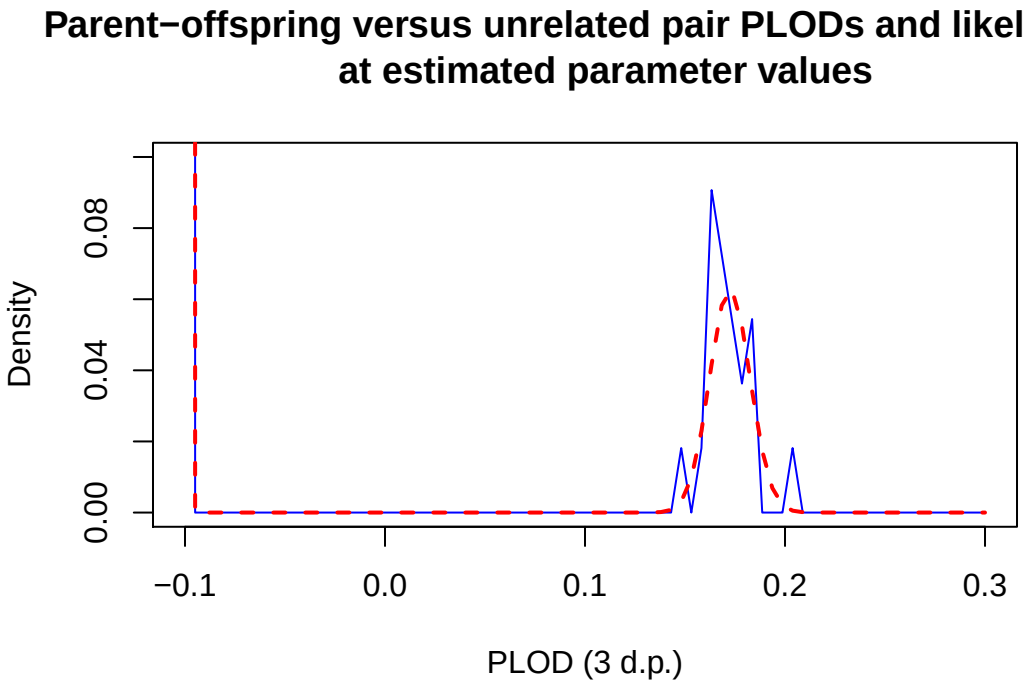


Figure 8.4: Distribution of PLODs for parent-offspring pairs within one sample.

Figures 8.1 to 8.3 demonstrate that the pseudo log-likelihood has minima at values of the parameters close to the true values. When N and ϕ are both allowed to vary there often seems to be some difficulty distinguishing them, not just in this particular simulation, with the likelihood frequently becoming bi-modal, with modes at higher and lower values of both parameters. This may cause lower precision in this preliminary model, and might be overcome when other kinships are included.

Figures 8.4 and 8.5 show examples of the observed distributions, and the pseudo-likelihood for the estimated parameter values, for the upper tail of PLODs for pairs within one sample, and between two samples, from the sample history above. The solid blue lines are the observed distributions and the dashed red lines are the likelihoods.

The observed PLODs are rounded to three decimal places, the saddlepoint approximations of their probabilities given the relevant kinships are calculated at those values, and the likelihood is calculated by weighting those approximations by the relevant kinpair probabilities given the estimated parameter values. As mentioned the distribution of PLODs for non close-kin pairs does not affect the estimates and is not computed. Instead, we assign to each such pair a point mass corresponding to the lowest value of the distributions plotted in Figures 8.4 and 8.5, as described in Section 8.1. The resulting distribution and probability mass function are used to find the maximum likelihood estimates.

To evaluate the usefulness of our preliminary method we simulated ten population histories as described in Section 7.4 and five sample histories for each. We used fewer sample histories here because fitting the model takes longer than finding the numbers of kinpairs. We also

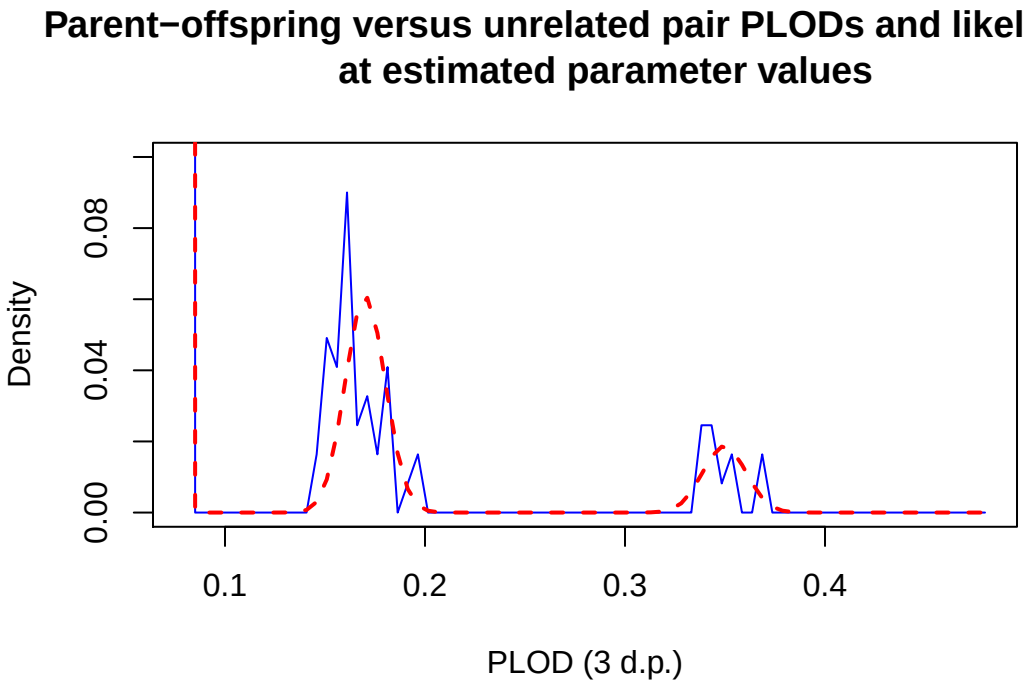


Figure 8.5: Distribution of PLODs for parent-offspring and self-pairs between two samples.

took much smaller, realistically sized samples of one tenth of the population at each survey (100 ~ 200 animals). The maximum likelihood estimates and coefficients of variation are presented in Figures 8.6 - 8.8.

The coefficients of variation and proximity of the mean estimates to the true values show that the estimates are useful for N, precise for ϕ, and very precise for λ, at realistic sample sizes. This confirms that our new method is promising for real applications. One caveat to repeat is that we used the historical breeding population allele frequencies from the simulation to approximate allele probabilities for each animal, whereas in reality we will have to use the sample allele frequencies. Also, producing a genotype of 1000 independent SNPs as simulated here is not trivial. On the other hand, in this proof-of-concept model we have not included PLODs between samples separated by more than the age of maturity, because our expression for the parent-offspring kinpair probability does not apply to those PLODs. However, our expression for the self-pair probability does apply to those PLODs, so the power of our estimates can likely be improved by including PLODs between samples separated by longer than the age of maturity, and including another term in our pseudo-likelihood which only distinguishes self-pairs from non-self-pairs. In real applications, samples are also often taken at shorter intervals, sometimes a few years in a row, which would allow us to compare more samples within the age of maturity and generate more information about parent-offspring pairs without increasing the number of samples.

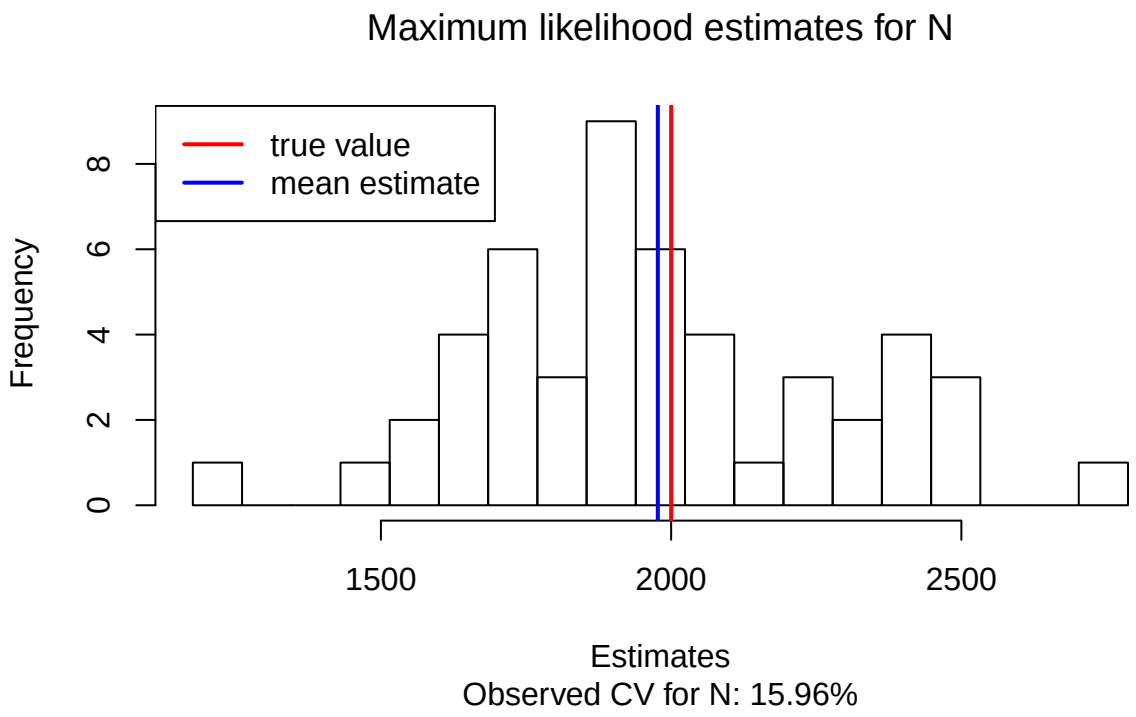


Figure 8.6: Estimates of N are distributed around the true value and the coefficient of variation is acceptable.

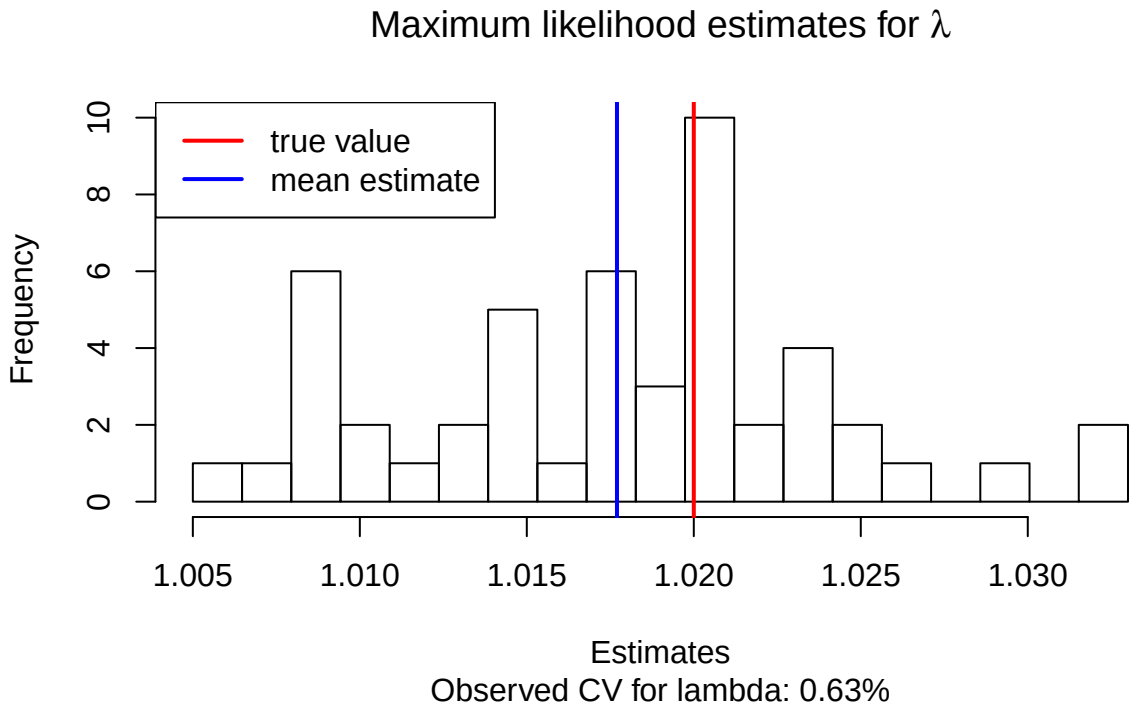


Figure 8.7: Estimates of λ are centered near the true value, and the coefficient of variation is very small.

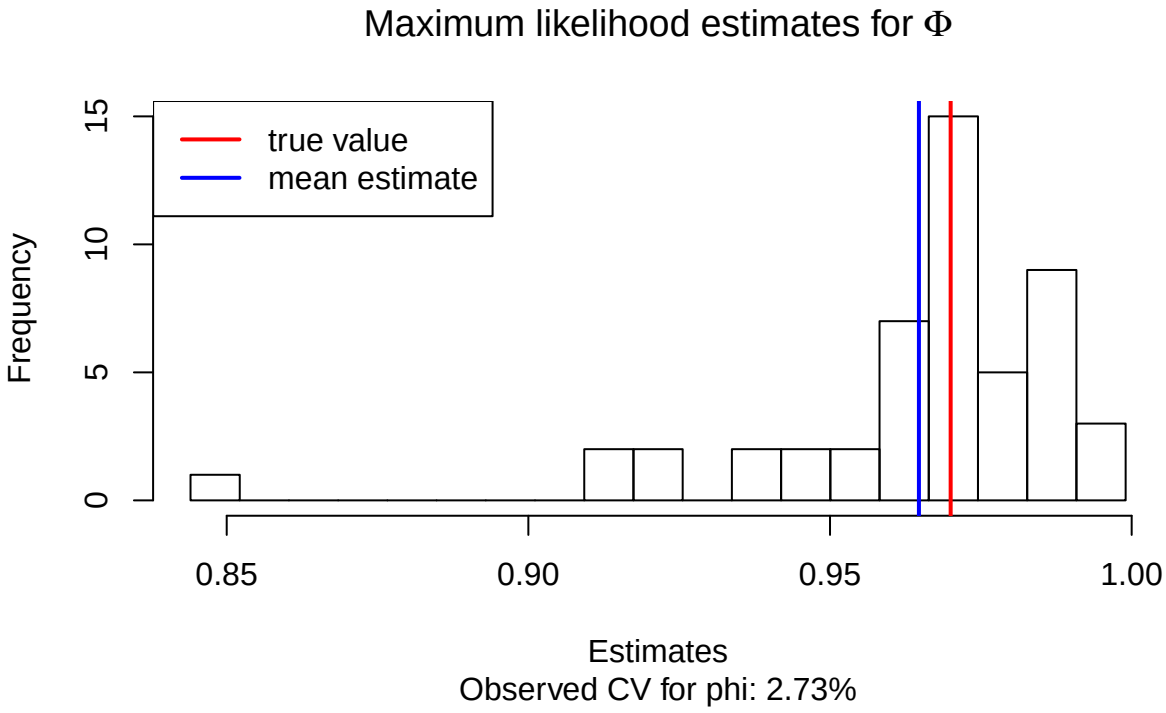


Figure 8.8: Estimates of ϕ are centered near the true value, and bounded at one, and there is an outlier at 0.85. The coefficient of variation is small.

Chapter 9

Summary

We have improved the speed of POPAN data simulation and model fitting substantially, which will enhance future releases and journal publication of the CAPOW capture-recapture power-analysis package. Future plans include migrating the new estimation code to the R package TMB, which should further improve the speed and stability of the optimization.

We have developed and validated a highly realistic simulation framework for the NZSRW population including genetic and pedigree information. This will be very helpful for investigating the interaction of demographics and population genetics in NZSRWs, and to produce data for developing new genetic modelling techniques.

We have described the calculation of close-kin pseudo log likelihood ratios (PLODs) in detail, and shown how their probability distributions can be described in terms of the kinships of the pairs of genotyped animals. This will be helpful as an introduction to the field of close-kin genetics, which is generating considerable interest, and for provoking new ideas in the field.

We have described how to apply the saddlepoint method to approximate the conditional distributions of close-kin pseudo log likelihood ratios given the kinships of the animals, and implemented the method and shown that it works in practice. These approximate densities open up possibilities for the development of many new statistical techniques using close-kin genetic data.

We have derived expressions for the probabilities of three close kinships for pairs of animals within and between sampling occasions, in terms of the demographic parameters corresponding to the size and growth rate of the population, and the annual survival probability of the animals. These expressions can themselves be used to estimate these parameters whenever numbers of kinpairs can be established, although our primary motivation is to incorporate them into a wider framework that absolves the analyst from making kinship decisions for each pair of genotypes.

We have combined our saddlepoint approximations and kinpair probabilities to form a preliminary pseudo likelihood for PLODs. We have shown how to implement maximum likelihood estimation of the demographic parameters based on this pseudo likelihood, and have

used our simulation framework to show that our approach has the potential to produce useful estimates at realistic sample sizes and parameter values.

There is much work that can be done to build upon this dissertation:

- CAPOW can be updated and re-released.
- The simulation framework can be streamlined by allowing validation data to be discarded while simulations are being run, and extended to allow user-input population and catch trajectories, life history traits, and genetic distributions, and to return the resulting datasets.
- The code for calculating PLODs, finding kinpairs in simulated data, generating saddlepoint approximations, calculating kinpair probabilities, and performing maximum likelihood estimation, can all be generalised and integrated into a Shiny interface or released as an R package.
- The bias introduced by approximating allele probabilities with sample allele frequencies can be quantified, and ways to mitigate it can be explored.
- The preliminary method can be applied to real-life data.
- Expressions for more kinship probabilities can be derived. The next most fundamental kinship is half-siblings, for which the calculation seems to be substantially more complicated than that for parent-offspring pairs, but together they may show how to express probabilities for all of the major kinships described in Section 6.3. We believe that this will significantly increase the power and applicability of our method.

Chapter 10

References

Andrews, C. (2010), The Hardy-Weinberg Principle. Nature Education Knowledge 3(10):65

Bravington, M. V., Skaug, H. J., Anderson, E. C. (2016), Close-Kin Mark-Recapture. Statist. Sci. 31, no. 2, 259–274. doi:10.1214/16-STS552. https://projecteuclid.org/euclid.ss/1464105042

Butler, R. W. (2007), Saddlepoint approximations with applications. Cambridge: Cambridge University Press.

Carroll, E. L., Childerhouse, S. J., Christie, M., Lavery, S., Patenaude, N., Alexander, A., Constantine, R., Steel, D., Boren, L. and Baker, C. S. (2012), Paternity assignment and demographic closure in the New Zealand southern right whale. Molecular Ecology, 21: 3960-3973. doi:10.1111/j.1365-294X.2012.05676.x

Carroll, E. L., Childerhouse, S. J., Fewster, R. M., Patenaude, N. J., Steel, D., Dunshea, G., Boren, L., and Baker, C. S. (2013), Accounting for female reproductive cycles in a super-population capture-recapture framework: application to southern right whales (*Eubalaena australis*). Ecological Applications 23, 1677-1690. https://doi.org/10.1890/12-1657.1

Carroll, E. L., Jackson, J. A., Paton, D., Smith, T. D. (2014), Two Intense Decades of 19th Century Whaling Precipitated Rapid Decline of Right Whales around New Zealand and East Australia. PLoS ONE 9(4): e93789. doi:10.1371/journal.pone.0093789

Carroll, E. L., M. Bruford, J. A. DeWoody, G. Leroy, A. Strand, L. Waits, J. Wang. (2018), Next-generation genetic monitoring using minimally invasive sampling methods. Invited Review for a Special Issue of Evolutionary Applications. 11: 1094-1119 DOI: 10.1111/eva.12600

Carroll, E. L., R. Alderman, J. L. Bannister, M. Bérubé, P. B. Best, L. Boren, C. S. Baker, R. Constantine, K. Findlay, R. Harcourt, L. Lemaire, P. J. Palsbøll, N. J. Patenaude, V. J. Rowntree, J. Seger, D. Steel, L. O. Valenzuela, M. Watson, and O. E. Gaggiotti. (2019), Incorporating non-equilibrium dynamics into demographic history inferences of a migratory marine species. Heredity. 122: 53-68

Charlton, C. M. (2017), Southern Right Whale (*Eubalaena australis*) Population Demographics in Southern Australia. Thesis presented for the Degree of Doctor of Philosophy, Curtin University. (http://hdl.handle.net/20.500.11937/59638)

Christiansen, F., Vivier, F., Charlton, C., Ward, R., Amerson, A., Burnell, S., et al. (2018), Maternal body size and condition determine calf growth rates in southern right whales. Mar. Ecol. Prog. Ser. 592, 267–281. doi: 10.3354/meps12522

Fewster, R. M., Jupp, P. E. Inference on population size in binomial detectability models, Biometrika, Volume 96, Issue 4, December 2009, Pages 805–820, https://doi.org/10.1093/biomet/asp051

Fewster, R. M. and Oh, J. (2015), CaPow – Capture-Recapture Power Analysis and model exploration. Version 1: EcoStats Symposium 2015. https://www.stat.auckland.ac.nz/~fewster/capow/

Harcourt, R., van der Hoop, J., Kraus, S., Carroll, E. L. (2019), Future directions in Eubalaena spp.: Comparative research to inform conservation. Front Mar Sci 530. https://doi.org/10.3389/fmars.2018.00530

Jackson, J. A., Carroll, E. L., Smith, T. D., Zerbini, A. N., Patenaude, N. J., Baker, C. S. (2016), An integrated approach to historical population assessment of the great whales: case of the New Zealand southern right whale. R. Soc. open sci.3: 150669. http://dx.doi.org/10.1098/rsos.150669

Jackson, J. A., Patenaude, N. J., Carroll, E. L., Baker, C. S. (2008), How many whales were there after whaling? Inference from contemporary mtDNA diversity. Mol. Ecol.17, 236–251. (doi:10.1111/j.1365-294X.2007.03497.x)

Hillary, R. M., Bravington, M. V., Patterso, T. A.n, Grewe, P., Bradford, R., Feutry, P., Gunasekera, R.,. Peddemors, V., Werry, J., Francis, M. P., Duffy, C. A. J., Bruce, B. D. (2018), Genetic relatedness reveals total population size of white sharks in eastern Australia and New Zealand. Scientific Reports volume 8, Article number: 2661 https://www.nature.com/articles/s41598-018-20593-w

International Whaling Commission [IWC] (2001), Report of the workshop on comprehensive assessment of right whales: a worldwide comparison. J. Cetacean Res. Manag. 2, 1–60.

Schwarz, C. J., and Arnason, A. N. (1996), A general methodology for the analysis of capture-recapture experiments in open populations. Biometrics 52, 860-873.

Xie, Yihui. (2015), Dynamic Documents with R and Knitr. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://yihui.name/knitr/.

Xie, Yihui. (2019), Bookdown: Authoring Books and Technical Documents with R Markdown. https://github.com/rstudio/bookdown.

Xie, Yihui. (2019), Knitr: A General-Purpose Package for Dynamic Report Generation in R. https://yihui.name/knitr/.

Zerbini, A., Ward, E., Kinas, P., Engel, H. M., Andriolo, A. (2011), A Bayesian Assessment of the conservation status of humpback whales (*Megaptera novaeangliae*) in the western Atlantic Ocean (Breeding Stock A). Journal of Cetacean Research and Management. Special Issue. 131-144.