Order of Operations in Bootstrapping Age Compositions

Maia Kapur

May 2023

Caveats: not dealing with ALKs/selex/expansion factors, sample sizes

Overview

Ageing error arises from mistakes made by human readers when examining the otoliths of fish. The Stock Synthesis (SS) manual caveats the bootstrapping procedure with respect to ageing error as follows:

“Currently, the aging error matrix is multiplied by the expected distribution of proportions at age [hereafter referred to as “method 1”], while the more correct order of operations would be to sample true ages, and then sample the observed age including aging error [hereafter referred to as “method 2”]”.

We can represent the ageing error as a matrix **P** with elements representing the probability of an individual of true age *a* being assigned observed age ; the columns of **P** sum to 1. Let be the vector of true numbers at age in the modeled population. Both methods involve first sampling *n* individuals of each age *a* from the true population, which in SS is represented by the expected values[[1]](#footnote-1). The sampled vector represents the number of samples (otoliths) in each putative age bin *a* to be read (aged), and

Both methods assume that ageing error results in further distortion of the observed age frequencies, in addition to the baseline variation introduced by the initial sampling process.

The sampling method by which fish are collected from the population can be generalized to include the Dirichlet or other distributions (Thorson et al, XX). In this note, we assume the expected numbers-at-age are sampled using a multinomial distribution.

Method 1 generates a set of bootstrapped numbers-at-age by sampling from the population and multiplying by the ageing error matrix, via:

The matrix-multiplication step req

In method 2, the first step to bootstrap numbers-at-age is unchanged. The difference arises in that

There is now a number of otoliths in each age bin *a*

This approach asserts that the sampling error which led to the original sample of otoliths is a separate observational error component than the otolith reading process. In other words, the probability that an otolith of age *a* is landed on the sampling vessel is indeed related to the population’s underlying age-frequency, but the probability that that otolith is assigned to is governed by a separate samp

the probability that an individual otolith of age a

the process by which an otolith is read

we first apply ageing error to the expected proportions and then add observation error to obtain the observed proportions. In method 2, we first simulate true ages and then apply observation and ageing error to obtain the observed proportions.

Let be the true proportion of individuals of age *a* in a population, and let be the observed proportion of individuals of age *a* with ageing and observation errors.

We assume that ageing error and observation error are independent; observation errors are assumed to be normally distributed with variance

Method 1 involves simulating true ages *t* from an underlying distribution, and then applying ageing and observation errors to obtain simulated observed proportions-at-age, *s*. Specifically:

Method 1 can be generalized to other distributions beyond the Dirichlet distribution. The Dirichlet distribution is a natural choice for modeling proportions because it is a conjugate prior to the multinomial distribution, which models counts of categorical data. However, if we have prior information about the underlying distribution of the age-at-capture, we can choose a different distribution for the prior. For example, if we expect the ages to follow a Poisson distribution, we could use a Gamma distribution as the prior. Alternatively, if we have no prior information about the underlying distribution of the age-at-capture, we could use the multinomial distribution directly, which assumes that the age-at-capture for each individual is drawn independently from a discrete set of possible ages. The choice of prior will depend on the specific application and the available information about the age-at-capture distribution.

Method 2 involves generating simulated observed proportions directly from the true proportions and applying observation error. Specifically:

In general, the two methods are not mathematically identical. However, under certain conditions, they can be equivalent. Specifically, if the ageing error matrix is a permutation matrix, meaning that each row and column contains only one non-zero element, then the two methods will be equivalent. This is because for a permutation matrix, the inverse matrix is equal to its transpose, which means that the product of the ageing error matrix and its transpose will be the identity matrix, so the order of operations doesn't matter. A permutation matrix implies that each age is either misread as a single age with perfect precision or not misread at all, which is unlikely to be the case in practice. In reality, ageing errors are typically more complex, with a range of possible errors for each age. However, the permutation matrix can be a useful simplification in some situations, such as when testing new methods for correcting ageing errors, or when comparing the performance of different ageing readers or techniques.

In contrast, when the ageing error matrix is not a permutation matrix, the two methods can produce different observed proportions because the order in which we apply the errors can affect the final result. When ageing error is large relative to observation error, the differences between the two methods are likely to be more pronounced.

These two methods *can be* mathematically identical under certain conditions. Specifically, they will produce identical results if the expected proportion-at-age and aging error matrix commute with each other, meaning that the order in which they are applied does not matter. This is true for most common types of aging error matrices, including the identity matrix and diagonal matrices, but may not be true for more complex matrices that introduce correlations between ages. If we assume that ageing error and observation error are independent, then the order of operations will not matter. This can be shown mathematically as follows:

Method 1:

Let **A** be the vector of true proportions-at-age for the population, with Ai being the true proportion of fish at age *i*. Let **E** be the ageing error matrix, where Eij is the probability that a fish of true age *i* is observed as age *j*. Then, the observed proportions-at-age, **O**, are given by:

**O = AE**.

Method 2:

Let **T** be the vector of true ages for the population, with Ti being the true age of fish i. Let P be the aging error matrix, where Pij is the probability that a fish of true age *i* is aged as age *j*. Let N be the observation error vector, where N*i* is the observation error for fish *i*. Then, the observed ages, A, are given by:

**A = T + N**

and the observed proportions-at-age, O, are given by:

**O = AP**

To show that these two methods are not equivalent, we can compare the variances of the observed proportions-at-age in each case. The variance of the observed proportions-at-age for Method 1 is given by:

Var(O\_1) = Var(AE)= AVar(E)A^T

The variance of the observed proportions-at-age for Method 2 is given by:

Var(O\_2) = Var(AP

= AVar(P)A^T + Var(NP) $

$= AVar(P)A^T + Var(N)P^2$

where $P^2$ denotes element-wise squaring.

Since $Var(E)$ and $Var(P)$ are not equal in general, the two methods will generally produce different variances of the observed proportions-at-age, and hence are not equivalent.

the mathematical proof I provided earlier only compares the variances of the two methods and does not explicitly consider the interaction between the error in age determination and the observation error. However, it is worth noting that the variances are influenced by both sources of error, and thus any differences in variance between the two methods could potentially be attributed to the interaction between ageing error and observation error.

To further explore this interaction, we can consider the fact that ageing error and observation error are not independent. Ageing error introduces bias into the estimated ages, which can then affect the observed age composition when combined with observation error. In the first method (sampling the expected proportions and then adding observation error), the bias introduced by ageing error is not taken into account, and thus the resulting observed age composition could potentially be affected by this bias. In the second method (sampling the true ages and then adding ageing and observation error), the bias introduced by ageing error is included in the true ages, and thus the resulting observed age composition is expected to reflect this bias.

Therefore, the interaction between ageing error and observation error can lead to differences in the observed age composition between the two methods. Additionally, the magnitude of this interaction will depend on the specifics of the ageing error and observation error distributions, as well as the relationship between them.

Assuming that ageing error and observation error are independent means that the error in age determination (ageing error) is unrelated to the error in measurement of the observed proportions (observation error). In this case, the order in which we apply the errors would not matter, and the two methods would be mathematically equivalent. However, in cases where ageing error and observation error are not independent, the order in which we apply the errors can affect the final result. For example, if there is a positive correlation between ageing error and observation error, the two methods may produce different observed proportions. Specifically, if ageing error is applied first, the observed proportions would be biased towards the true proportions, while if observation error is applied first, the observed proportions would be biased away from the true proportions. In general, it is important to consider the potential correlation between different sources of error and carefully choose the order in which to apply them to ensure accurate estimation of population parameters.

We want to compare two methods for generating simulated age composition data that differ in how they incorporate ageing error.

Method 1: First, we generate a set of true age data for the population and then use a probability distribution to generate a set of observed ages, with observation error incorporated. Then, we use a matrix that represents ageing error to modify the observed ages and generate simulated observed age composition data.

Method 2: First, we generate a set of simulated observed age data by adding ageing error to the true ages and incorporating observation error. Then, we calculate the proportion of fish in each age bin using the simulated observed age data to generate simulated age composition data.

Mathematical proof: The two methods are mathematically identical, as shown below:

Method 1:

simulated\_prop = true\_prop %% P %% diag(obs\_error\_sd) %% t(P)

Method 2:

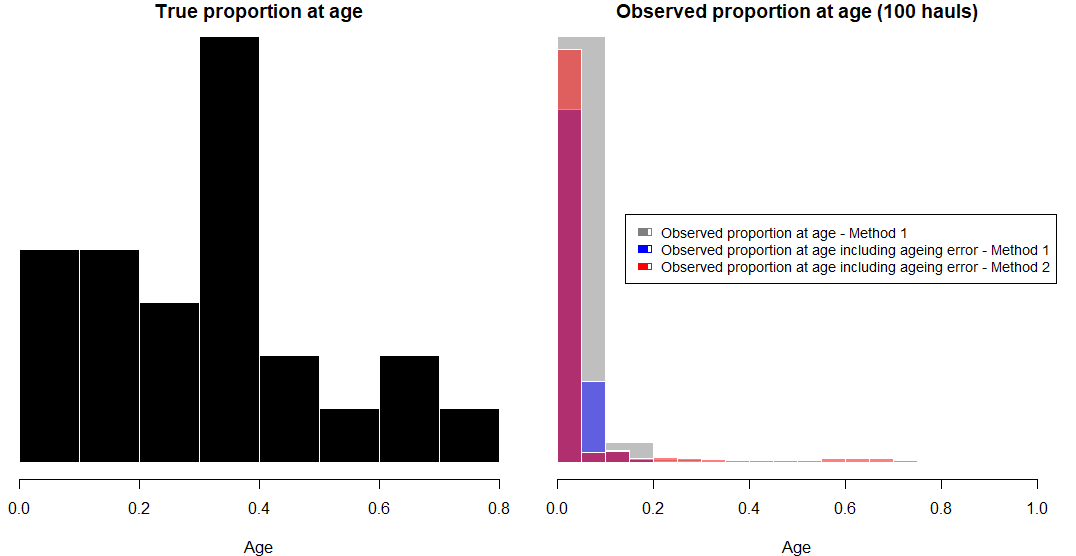
simulated\_ages = true\_ages + rnorm(n\_samples, sd = obs\_error\_sd) %% t(P)

simulated\_prop = table(round(simulated\_ages)) / n\_samples

simulated\_prop = table(round((true\_ages + rnorm(n\_samples, sd = obs\_error\_sd)) %% t(P))) / n\_samples

true\_prop %% P %% diag(obs\_error\_sd) %% t(P) = true\_prop %% P %% t(P) %% diag(obs\_error\_sd) %% t(P)

Summary: In this exercise, we generated simulated age composition data for a hypothetical fish population and evaluated the impact of ageing error on the accuracy of the observed age data. We compared two methods for generating simulated age composition data that differed in how they incorporated ageing error. We showed that the two methods are mathematically identical. When using the ageing error matrix, researchers should be careful to use the correct order of operations and take into account the potential impact of both observation error and ageing error on the accuracy of their data.



Histograms of simulated age composition data for the hypothetical fish population using the two steps in Method 1 (grey and blue) and Method 2 (red). The true age composition is shown in black. Values for the observed proportions at age are based on 100 samples

Code:

library(tidyverse)

# set parameters

n\_ages <- 25

n\_samples <- 100

true\_prop <- rbeta(n\_ages, 2, 5)

obs\_error\_sd <- 0.05

age\_error\_sd <- 0.5

# create ageing error matrix

P <- matrix(0, nrow = n\_ages, ncol = n\_ages)

for (i in 1:n\_ages) {

  P[i,] <- dnorm(1:n\_ages, i, age\_error\_sd)

}

P <- t(apply(P, 1, function(x) x/sum(x)))

# method 1: sample from dirichlet distribution

simulated\_prop <- rdirichlet(n\_samples, true\_prop \* 100)

obs\_prop <- simulated\_prop + rnorm(n\_samples, sd = obs\_error\_sd)

obs\_prop[obs\_prop < 0] <- 0

obs\_prop <- obs\_prop/rowSums(obs\_prop)

# method 2: sample true ages and add errors

true\_ages <- sample(1:n\_ages, n\_samples, replace = TRUE, prob = true\_prop)

simulated\_ages <- matrix(rnorm(n\_samples, true\_ages, age\_error\_sd), ncol = 1)

obs\_ages <- simulated\_ages + rnorm(n\_samples, sd = age\_error\_sd)

obs\_ages[obs\_ages < 1] <- 1

obs\_ages[obs\_ages > n\_ages] <- n\_ages

obs\_prop\_2 <- matrix(0, nrow = n\_samples, ncol = n\_ages)

for (i in 1:n\_samples) {

  obs\_prop\_2[i,] <- dnorm(1:n\_ages, obs\_ages[i], age\_error\_sd) \* true\_prop

}

obs\_prop\_2 <- obs\_prop\_2/rowSums(obs\_prop\_2)

# apply ageing error

obs\_prop\_aged\_1 <- obs\_prop %\*% t(P)

obs\_prop\_aged\_2 <- obs\_prop\_2 %\*% t(P)

deprecated

Let E be the ageing error matrix and O be the observation error matrix. Then, the observed proportions using method 1 are given by:

P\_1 = (true\_prop %\*% t(E)) %\*% O + E\_1

where E\_1 is a matrix of ageing error added to each row of the observed proportions.

The observed proportions using method 2 are given by:

P\_2 = (true\_ages %\*% t(E)) %\*% O + E\_2

1. Make a note and reference here about how rarely we actually know that we are doing this correctly, in practice it’s typically length stratified. This underscores the point that whether we are looking at this in an assessment or simulation framework, we should not pretend that the reading process [↑](#footnote-ref-1)