

Lab Assignment 3

STA 100 | A. Farris | Spring 2021

*A pdf copy of your assignment is due at 5pm PT Monday, May 3. Submission of the pdf will be through Gradescope. Please put in the effort to make it look reasonably professional – you’re encouraged to use R Markdown. Note that specific tasks for you are **highlighted**. You are free to work in groups, but you must submit your own writeup, and run your own code.*

Introduction to mark-recapture

How large is a population of Humpback whales? This is no trivial question for researchers: one can’t simply go out and count them!

One study attempted to answer this question by photographing whales in a breeding area in two consecutive years. After the first year, the photographs were closely compared, and unique whales identified.

After the second year, the photographs were again closely compared, and unique whales identified; moreover, photos from the first and second years were closely compared, so that whales seen more than once could be identified.

Roughly speaking, if the proportion of the whales seen in the first year that are seen again in the second were to be very high, then it seems that the total number of unique whales in the population should be not much larger than the number of whales seen in that second year. On the other hand, if the proportion of the first years’ whales seen again were to be small, then the total number of unique whales in the population seemingly should be much larger than the number seen in the second year.

The name mark-recapture for this methodology comes from the idea of sampling from a population, marking the sampled individuals, returning them to the population, resampling from the population, and then checking how many of the ‘marked’ individuals have been ‘recaptured.’

A probability model

In order to be somewhat more precise with this reasoning, let’s begin by making some simplifying assumptions. Firstly, let’s assume that the population doesn’t change from year to year: the whales present in the breeding area one year are present again the next.

Secondly, let’s assume that the photographed whales are randomly sampled from the population. Because unique whales are identified, we can presume that the sampling is carried out without replacement.

Let’s assume further that the total number of whales in the population is N , the number of unique whales seen in the first year is M , and the total number of unique whales making up the sample seen in the second year is n . This situation is equivalent to drawing n things randomly *without* replacement (the whales are unique, hence not counted more than once) from a population of size N , M of which are distinguished in some way. The distribution of the number of the distinguished things in the sample is called *hypergeometric* with parameters N , M , and n .

Thus, given the numbers N , M , and n , we could assign specific probabilities $P(X = m)$ to events in which possible numbers of whales m are seen again, given by a hypergeometric distribution.

However: at the end of the study, we will know the numbers M and n of unique whales seen in the two years. We will also observe one of these possible values m of whales seen again the second year to have occurred. We will still not know the total number N of whales in the population, though.

The hypergeometric distribution will assign specific probabilities to hypothetical values of m for given values of N , M , and n . How could we use this to go about *estimating* N from M , n , and the observed m ?

The Likelihood approach

The likelihood approach to a problem such as this simply suggests that, roughly speaking, the possible parameters be judged according to the probabilities that they would hypothetically assign to the observed value. When these probabilities are thought of as a function of the unknown parameter value, we call them *likelihoods*.

We call these likelihoods rather than probabilities because these may be considered after the sampling has taken place, when the random number X has been observed to take some value m . Strictly speaking, then, it might no longer make sense to refer to these probabilities expressing chances with which events will happen.

The possible parameter assigning the highest likelihood can be chosen as the *maximum likelihood estimate* (MLE) of the unknown parameter, under our probability model and with the observed data.

Estimating the population size

It is possible to investigate likelihoods with many probability models analytically, but here we will estimate MLEs for the unknown population size numerically (that is, using computers to do calculations for us).

For an example, let's say that 10 unique whales were seen the first year, 20 were seen the second, and that 4 of the whales were seen again the second year. In this case four out of the ten whales seen the first year were seen again, which is not a large proportion; so it seems that the population size should be substantially larger than twenty, the number seen in the second year.

To evaluate this using the likelihood approach, we can begin by defining a function `hyperLik` that will calculate likelihoods, based on different values of the numbers N , M , n , and m .

```
hyperLik <- function(N,M,n,m){dhyper(m,M,N-M,n)}
```

We can use this for example to say that the likelihood assigned to a population size of $N = 50$ would be

```
hyperLik(50,10,20,4)
```

```
[1] 0.2800586
```

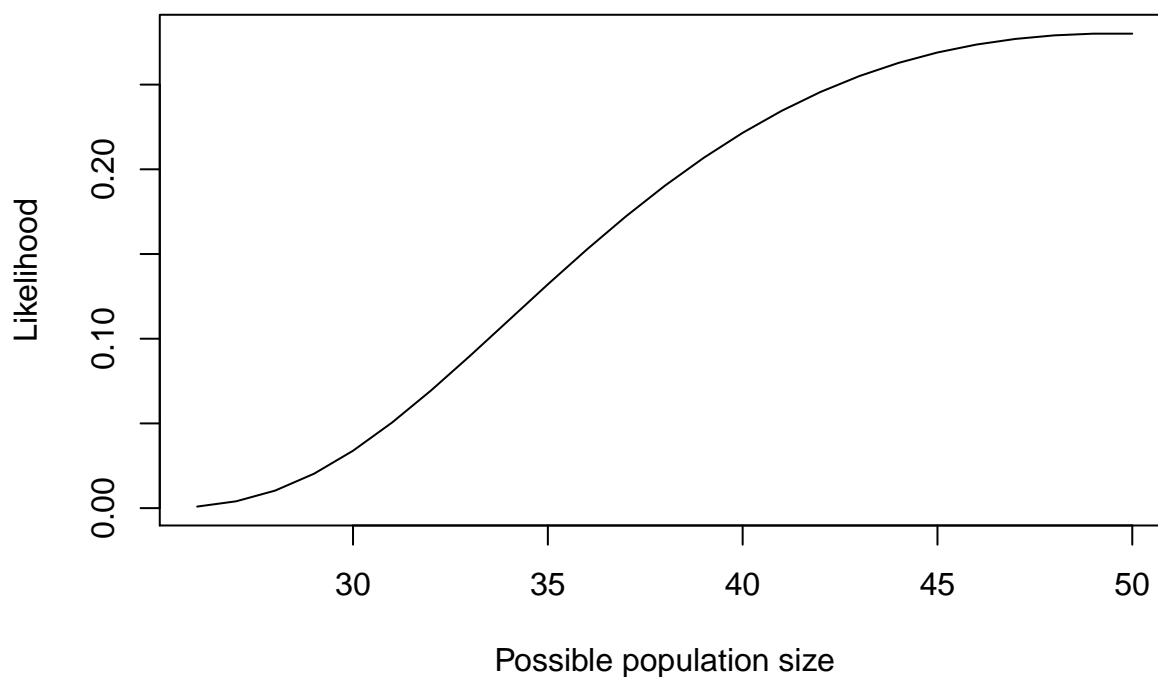
while the likelihood assigned to a population size of $N = 100$ would be

```
hyperLik(100,10,20,4)
```

```
[1] 0.0841073
```

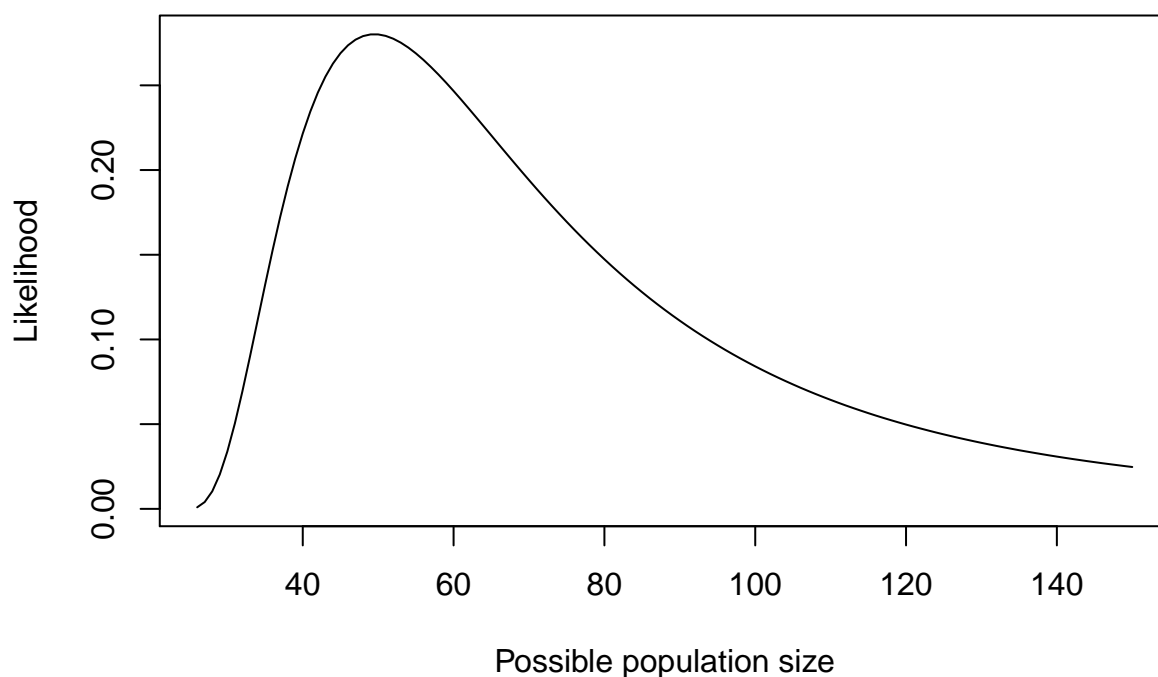
suggesting the former possibility to (roughly speaking) better fit the data. In this case we could find the possible value of N maximizing the likelihood by computing the likelihood for many of the possible values of N . The smallest possible value of N here is the number of whales seen in the second sample, 20, plus the number from the first sample that were not seen again (which is $10-4=6$), i.e. $20+6=26$. In principle, any population size larger than 26 is possible. We can compute the likelihoods for possible values from 26 to 50, for example, and plot them:

```
possibleN <- 26:50
likelihoods <- hyperLik(possibleN,10,20,4)
plot(possibleN,likelihoods,
     type="l", # linear interpolation for a smooth curve
     ylab="Likelihood",
     xlab="Possible population size")
```



Seeing that values close to 50 seem to be most likely, we should look at a larger range of possible values to be sure of seeing the highest likelihood. Looking instead at values from 26 to 150:

```
possibleN <- 26:150
likelihoods <- hyperLik(possibleN,10,20,4)
plot(possibleN,likelihoods,
     type="l",
     ylab="Likelihood",
     xlab="Possible population size")
```



With this larger range we can be more confident of seeing the most likely values, which appear to be near 50. We can find an exact value with the highest likelihood by

```
est <- possibleN[which.max(likelihoods)]
est
```

Note that in the event of ties, this gives us the smallest population size with the maximal likelihood. In this case our estimate for the population size would be 49.

Using photos, researchers identified 1,377 and 1,467 individual whales in a breeding area during the first and second years of their study, respectively. Of those that they identified in the second year, 316 had been seen the previous year as well. Is the proportion of the whales seen in the first year that are seen again in the second large? What does this indicate about the size of the population relative to the size of the sample in the second year? Plot the likelihoods for possible values of the population size. What would you estimate the population size to be?

In a separate study, researchers captured and tagged 16 grizzly bears. After the tagging, wildlife cameras in the area saw 11 different bears, of which 8 were tagged. Is the proportion of tagged bears that were seen by the cameras large? What does this indicate about the size of the population relative to the size of the sample seen by the cameras? Plot the likelihoods for possible values of the population size. What would you estimate the population size to be?

Appendix: R Script

```
hyperLik <- function(N,M,n,m){dhyper(m,M,N-M,n)}
hyperLik(50,10,20,4)
hyperLik(100,10,20,4)
possibleN <- 26:50
likelihoods <- hyperLik(possibleN,10,20,4)
plot(possibleN,likelihoods,
     type="l", # linear interpolation for a smooth curve
     ylab="Likelihood",
     xlab="Possible population size")
possibleN <- 26:150
likelihoods <- hyperLik(possibleN,10,20,4)
plot(possibleN,likelihoods,
     type="l",
     ylab="Likelihood",
     xlab="Possible population size")
est <- possibleN[which.max(likelihoods)]
est
```