STA 138 Discussion 4 Solutions

Fall 2020

Data analysis

For our discussion this week, we will explore the use of the maximum likelihood approach to estimation in a mark-recapture setting.

A study set out to estimate the size of a population of humpback whales. To do so, extensive aerial photographs were taken of whales in a breeding area in each of two consecutive years. After the first year, the photographs were closely compared, and unique whales identified (each was assigned an ID number). After the second year, then, 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 in both years could be identified.

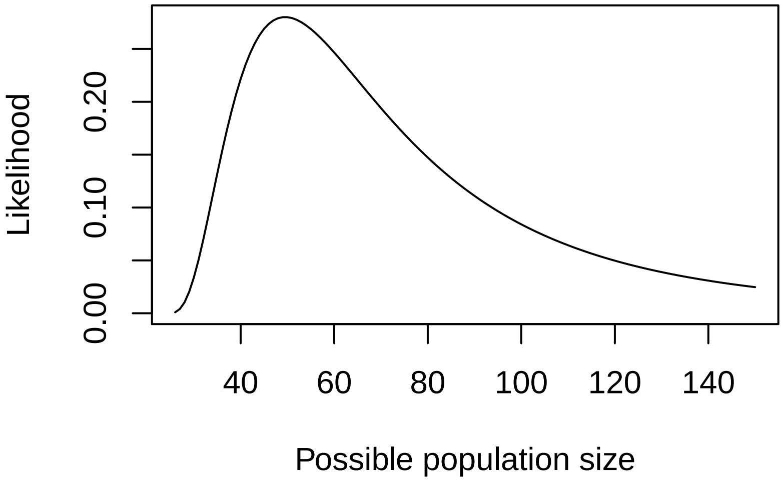
We will assume that the whales that were seen were randomly sampled from the population in each year, and that the populations in the two years were the same.

Suppose that 10 unique whales were seen the first year, 20 were seen the second, and that 4 of thse whales were seen again the second year (i.e. seen in both years). N=?, M=10, n=20, k=4

(a) Plot the graph of the likelihood function given the results above.

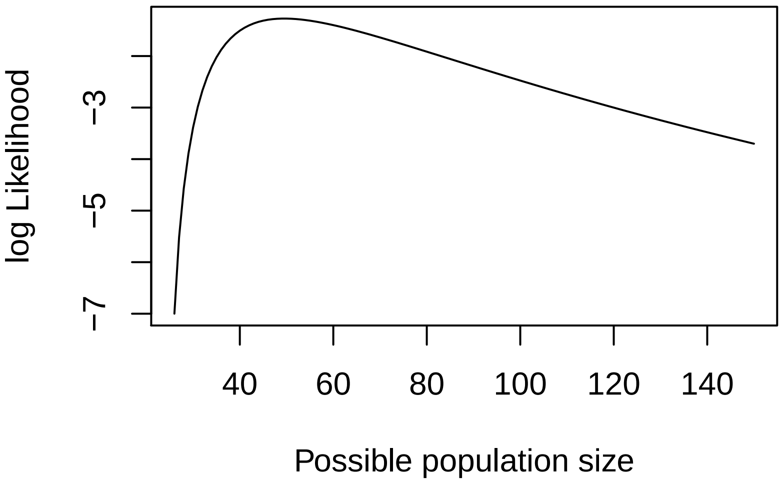
To do this, we can simply compute the likelihood at each possible value of the parameter within some sufficiently large range. It is necessary, though, to identify what values are possible. The smallest possible value of N here is the number of whales seen in the second sample, 20, plus the number from the first sample that were not seen again (which is 10-4=6), i.e. 20+6=26. 10+20-4=26

We also need to be careful to plot this over a sufficiently large range of values. In optimizing complicated functions, we should be wary that local maxima are not always global maxima (though in this case our likelihood function is fairly well behaved)!



(b) Plot the log-likelihood function given the results above.

Here we simply plot the logs of the likelihoods obtained above.



(c) What is the maximum likelihood estimator of the population size on the basis of these results? Do the likelihood and log-likelihood in (a) and (b) yield different maximizers? Explain.

To maximize the functions above over possible integers N, we can use a simple grid search. This is feasible in this case, although we should note that in more computationally demanding cases we might make use of more sophisticated numerical methods.

The candidate value maximizing the likelihoods here is 49. Both the likelihood and log likelihood admit the same maximizers, by necessity, because the natural logarithm is a monotone increasing function!

Code Appendix

```r
hyperLik <- function(N,M,n,m){dhyper(m,M,N-M,n)} # reparameterizing to N, M, n
possibleN <- 26:150
likelihoods <- hyperLik(possibleN,10,20,4)
plot(possibleN,likelihoods,
     type="l",
     ylab="Likelihood",
     xlab="Possible population size")
plot(possibleN,log(likelihoods),
     type="l",
     ylab="log Likelihood",
     xlab="Possible population size")
# MLE:
est <- possibleN[which.max(likelihoods)]
# note that in event of ties, this selects the smaller value!
```