Lab 1 - Setting the Stage

Erik Blomberg (edited by Matt Mensinger and Liam Berigan)

07/07/2023

Contents

1 Lab Overview **2**

1.1 Learning objectives 2

2 The Lincoln-Petersen Estimator **2**

3 Lab Setup - Estimating Abundance **4**

3.1 Simulating a mark-recapture sample 4

3.2 Spreadsheet basics 6

4 Completing the Lab Exercise **7**

5 Try it in R **9**

6 The Three Big Things You Should Have Learned Today **10**

7 Lab 1 Assignment **10**

8 Lab Appendix **12**

8.1 Quick reference for important R commands in this lab 12

8.2 For more information 12

8.3 Glossary of Terms 13

1 Lab Overview

In this week's lab we will review some old concepts and set the stage for the next 4 weeks of class where we will be using a combination of Excel and R to explore a variety of concepts in population ecology and demographic estimation.

1.1 Learning objectives

1) Provide recap of some important concepts from past classes, including basics of mark-recapture and its assumptions, and some statistical concepts like basic principles of mean, variance, SD, and confidence intervals.

2) A refresher on working in Microsoft Excel and an introduction to the expectations for this class around building Excel spreadsheets and using them to execute mathematical models via linked formula.

3) Continue to get your feet wet in the realm of coding in R and writing scripts to execute some of the same procedures you complete in Excel.

2 The Lincoln-Petersen Estimator

You should recall that the Lincoln-Petersen equations allow us to derive estimates of abundance based on a very simple mark-recapture study design. The estimators are based on a simple ratio of marked to unmarked individuals, where some number of animals are captured and marked, and then during a second sampling period another group of animals is captured and the number of marked vs. unmarked animals are recorded. This is referred to as a **batch marking** process, where individuals are not uniquely identified. Or rather, they could be **individually marked**, but we do not use the individual identities as part of our estimation.

We assume the population is closed between the two samples: no animals are born, die, or move away from the sampled area. That is to say that when using a Lincoln-Petersen equation we assume that every individual marked in the first sample is available to be detected during the second sample. This also means that we must be able to recognize and record the presence of marks on the animals, and that no marks are lost. Given all these assumptions, the Lincoln-Petersen estimator is referred to a **closed capture model** - we

will learn more about **open capture models**, which accommodate loss of marked animals, later in the semester.

The Lincoln-Petersen equation for estimating abundance is given as

$$\frac{\hat{N}}{M} = \frac{C}{m} \tag{1}$$

Where $\hat{N}$ is the estimate of total abundance, M is the number of animals marked in the first sample, C is the total number captured in the second sample, and m is the number of animals captured in sample 2 that were previously marked. This formula is intuitive; the ratio of total population size to the number of marked individuals in the original sample should be the same as the ratio of the size of the second sample to the number of marked individuals in the second sample. If we want to solve for $\hat{N}$ (which we normally do) then we can re-arrange the equation as follows:

$$\hat{N} = \frac{CM}{m} \tag{2}$$

It turns out that the Lincoln-Petersen estimator can become **asymptotically biased** at small sample sizes, especially when there are few recaptures of marked animals. This is because small values of m in the denominator of the equation will tend to produce estimates of $\hat{N}$ that are biased large. For this reason it is common to use the following "un-biased" formula as an alternative:

$$\hat{N} = \frac{(C+1)(M+1)}{m+1} - 1 \tag{3}$$

Notice that in this case as the values for C, M, or m get increasingly large, the influence of adding a value of 1 to each becomes increasingly small. So, even if your sample size is large it is probably prudent to use the bias-corrected estimator.

We will typically also want to calculate a measure of variance, standard deviation (SD), and/or confidence intervals for our estimate of $\hat{N}$. An approximate estimate of the variance of $\hat{N}$ can be derived using the following equation:

$$Var(\hat{N}) = \frac{(C+1)(M+1)(M-m)(C-m)}{(m+1)^2(m+2)} - 1 \tag{4}$$

During the lab today, you'll use this estimate of variance to calculate two other measures of precision of an estimator; its Standard Deviation (SD) and its 95% Confidence Interval.

These should be familiar concepts from past courses, such as WLE 220, but just as a refresher, recall that we can calculate the SD of the variance as $SD{=}\sqrt{Var}$, and the 95% Confidence intervals are approximated as $SD * 1.96$, reflecting the fact that for a normal distribution, 95% of the observations are said to be within 1.96 standard deviations of the mean. The 95% Confidence Interval is bounded by the 95% Confidence Limits, where the lower limit occurs at $\bar{X} - (SD * 1.96)$, and the upper limit occurs at $\bar{X} + (SD * 1.96)$.

3 Lab Setup - Estimating Abundance

For this lab, we'll be working in a combination of Microsoft Excel and RStudio, where we will use R/RStudio to generate some simulated mark-recapture data, and we will use Excel to build our Lincoln-Peterson estimators and execute the equations.

3.1 Simulating a mark-recapture sample

During normal, in-person labs, we would accomplish the following data collection using dried beans and sharpie markers. I would give you a 'population' of beans which you would sample from, apply marks, resample, and use the resulting data as your values for C, M, and m. In our virtual version of this lab, we will use R to simulate recapture data for us. Not nearly as fun as chasing rogue pinto beans around the room, but just as illustrative. And I'll try to make it at least mildly entertaining.

First, because dealing with abstract numbers is no fun, let's choose a study organism. Run the code below and return the 'Species' object to get your study organism.

```
Species <- sample(c("White-tailed Deer", "Wild Hog", "Wooly Mammoth",
    "White-footed Mouse", "Wallaby", "Wandering Albatross", "White Pelican",
    "Wilson's Warbler", "Whooping Crane", "Wood Duck", "Wood Frog",
    "Whiptail", "Western Pond Turtle", "Western Diamondback",
    "Wood Turtle", "White Bass", "Whale Shark", "Wahoo", "Walleye",
    "Wobbegong"), 1)

Species
```

Next, we will generate some simulated mark recapture data using a very simple series of commands. Run the following code:

```
# this will return your true population size (spoiler
# alert)

N.True <- rpois(1, 400)

# return that value

N.True

# number captured and marked in the first sample

M <- 40

# number captured total in the second sample

C <- 40

# this will return the number of previously marked animals
# in the second capture

m <- rbinom(1, C, prob = (M/N.True))

# return that value

m
```

Sidebar - the anatomy of this code - It's not totally necessary for you to understand this code for the exercise to work, but if you're interested, the command rpois() returns a randomly selected whole number with a mean expected value of 400. The 1 in the parenthesis tells R to do it 1 time. The rbinom() command is similar, but it says for the 40 individuals that were recaptured during the second sample, randomly choose how many of them are caught again in the second sample based on a probability of capture equal to M (the number marked) divided by N (the total population size), which simply indicates that if there are 40 marked animals in a population of 200 total animals, the probability that any one animal you catch will be marked is equal to 40/200 = about 0.20. If we can

A	B	C
1	3	=A2+B2
4	2	=A3+B3
6	1	=A4+B4
2	7	=A5+B5

meet the assumptions we discussed at the beginning of the lab, then it stands to reason that about 20% of our C=40 animals in the second sample will also be marked. Overall, the rpois() and rbinom() commands introduce some random variation (that is what the 'r' stands for) in the sampling, such that none of us should get exactly the same numbers from this exercise.

Open Microsoft Excel. Create a spreadsheet to house your values for N, M, C and m (they should all be printed in your R console). Before you get too far in the structure of this spreadsheet, see the following guidance:

3.2 Spreadsheet basics

Before you get started, I have some guidelines I would like you to follow for using spreadsheets in this and future classes.

1) Your spreadsheet should conform to the rectangular layout shown in Figure 4 of Broman and Woo's excellent 2017 paper on spreadsheet basics, which can be found at this link and is also posted as a pdf on Brightspace under the 'Readings and Resources' page. See Figure 5 of their paper for a bunch of examples of what not to do.

2) All calculations should be made via formulas entered directly into cells in your spreadsheet. When you are making calculations based on other values contained within the spreadsheet, you should link the formulas to those cells directly. For example, if I had two values of A and B, and I wanted to calculate A+B=C in my spreadsheet, it should look like the following:

3) Only data, calculations, and labels belong in your spreadsheets. I know in past classes you've been include other components, such as a listing of your formulas, in excel files in order to show your work. For this class, we will consider spreadsheets to be tools that you use to store and analyze data, and the embedded formulas you use for making calculations will constitute your work being shown.

A	B	C
X	4	Do This
X=4		Not This

4) One cell, one piece of information. This means that if you have a label, and a value, they should go into two different cells. This is particularly important for numeric values, because if you include a symbol or character in a cell with a number, Excel will no longer read the number and you will not be able to apply formula to it. For example:

5) Be MINDFUL of the order of operations when meshing the equations provided above to the formula you enter in Excel. Most mistakes happen because of misplaced or too few parentheses. Also keep in mind that in Excel, you will need to place an * to indicate multiplication - i.e. - if the formula says (C+1)(M+1), your formula should include (C+1)*(M+1). Also as a reminder, you can use a 'carrot' ^ to indicate raising something to a power - i.e. X^2 is the same as X^2.

4 Completing the Lab Exercise

We will generate a few more simulated datasets to hopefully illustrate how estimates of abundance change as we alter certain aspects of the data, such as sample sizes. To complete the lab, you will use your Excel table to estimate $\hat{N}$ using ***equation 3*** and its variance using ***equation 4***. From these, you will derive the SD and the 95% Confidence Interval of $\hat{N}$.You should repeat this for each of the scenarios described below. For full clarity, your order of operations for completing each scenario is:

1) Modify the R code provided above to generate data for the particular scenario

2) Enter the new values for N.True, M, C, and m into your spreadsheet.

3) Use your existing linked formula in Excel to generate estimates of $\hat{N}$, Var, SD, and 95% CI for the new scenario.

Sidebar - A reminder on model object naming - Remember, each time you run a line of code to create an object of a particular name, it will overwrite any previous objects with that same name. So as you go through these excercises,

you should record values for M, C, and m in your spreadsheet, along with any relevant labeling so you can keep the iterations organized and make sure you have the correct values for each. Alternatively, you could write your R script so that you create new, different obects for each scenario. For example, instead of running M<-80 for scenario 1, you could run M.S1<-80, using M.S1 to reflect it is the value of M for scenario 1. But recognize you would then need to use the object M.S1 in any subsequent code for that scenario where just M was used previously. You're not required to customize your code this way - it's just a suggestion.

Scenarios to complete:

1) *The OG* - this is the scenario based on the initial values above with C=40 and M=40. You already have this one done!

2) *Greater sample size* - Increase the number of animals marked in the first sample (M) to 80, and then re-sample (C) 80 animals. In doing so, recognize you are sampling a larger segment of the population by virtue of your increased sample size.

3) *Even GREATER sample size* - Repeat step 1, but dial the number up to 120 for both M and C. Thus we will have 3 sample sizes (40, 80, and 120) to evaluate the effect of sample size on our population estimates and their precision (SD, 95% CI).

4) *Introducing error* - Using the same values of N.True, M, and C from scenario 2, but replace the code for m with the following:

```
m <- rbinom(1, C, prob = (M * 0.8/N.True))

m
```

This is the same code we used above, but with a small change to multiple the value of M by 0.8. This will simulate a situation where of the 80 animals marked in the first sampling occasion, only 80% remain available in the second occasion. This would be analogous to violating an assumption of closure, for example, if a proportion of the population left the study area or died between the first and second sample.

5) *Future Scenario* - Assume that you conduct a second mark-recapture study of your animal population 5 years after the first sampling event. In this case you mark 100

beans and recapture 100, where 23 of the second sample are marked. What are the estimates of $\hat{N}$ (with Var, SD, and confidence intervals) for the bean population 5 years later?

Using your estimate of $\hat{N}$ from Scenario 1 as your starting abundance, and principles of exponential growth that we have covered in lecture, what was the cumulative rate of change (λ_5) during the 5-year interval, and what was the annual rate of change (λ)? Given the value for annual λ that you calculate, what do you predict will be the bean abundance in year 8? How long will it take for the population to double in size AFTER year 5? Incorporate these answers into the word document that you will turn in as part of your assignment. I recommend adding also adding these formulas to your Excel spreadsheet, even if you worked out the answer on paper. This will allow us to see where you went wrong if you get the wrong answer.

5 Try it in R

I decided to keep a major portion of this and the next 3 labs in Microsoft Excel, because I think there is great value in you becoming more proficient in Excel and being able to use it for building relatively simple ecological models. But, everything we do in this class can be replicated in R, and the Lincoln-Petersen Model is no different. To illustrate this, let me revisit our very first data with M and C of 40, and I'll work through generating an estimate of $\hat{N}$ using ***equation 2***. I would like you to modify the equation to produce the unbiased estimator provided by ***equation 3***. You will know it's working if it returns an estimate of $\hat{N}$ somewhere in the ballpark of the true estimate. Leave your code modification in your R script; there is no need to do anything extra for the assignment.

```
# uses the same value of N.True you've been working with
# all along

# number captured and marked in the first sample

M.R <- 80

# number captured total in the second sample

C.R <- 80
```

```r
# this will return the number previously marked in second
# sample

m.R <- rbinom(1, C.R, prob = (M.R/N.True))

# estimate of n-hat based on equation 2

N.hat.R <- (C.R * M.R)/m.R

# return the estimate

N.hat.R
```

6 The Three Big Things You Should Have Learned Today

FIRST Today should have served to remind you about principles of variance, standard deviation, and confidence intervals, and how they all related to one another. We will continue to use these throughout the semester.

SECOND You should have picked up some best practices for building spreadsheets in Microsoft Excel, including translating mathematical equations into Excel formulas that are linked to data contained in the spreadsheet.

THIRD The scenarios should have enforced the role of sample size in generating robust population estimates, and the consequences of violating assumptions.

7 Lab 1 Assignment

For this assignment, turn in your completed spreadsheet with correctly obtained estimates of $\hat{N}$, its variance, SD, and 95% CIs for each of the scenarios listed above. Also include your calculations and future predictions on this spreadsheet, but you do not need to compute variance, SD, or 95% CIs for the year 8 estimate. Produce a figure that shows all $\hat{N}$ estimates (40, 80, error, and future predictions), the confidence intervals (where appropriate), and illustrates the true population size.

NOTE – Students often struggle when adding error bars to their figures. I recommend using Google if you need help learning and/or remembering how to use this feature of Excel.

Your assignment will be to turn in the Excel Spreadsheet (i.e. the actual digital file) with your completed calculations and figure. You may leave the figure embedded in the Excel file and there is no need to write a figure caption to go with it. Format for figures should conform to the follow guidelines, which generally follow those of the Journal of Wildlife Management:

All text should be in Arial 12 pt font (note that Calibri is the default and is not acceptable). The figure should be presented in black and white only and all grid lines should be removed. Axis titles may be in bold, but all other text in standard font style. Capitalize the first word of the axis title with all other words in lower case unless a proper noun. Remove all outer borders from the figure and the legend. There should be no figure title.

In addition to the Excel spreadsheet, please provide a short (~1 paragraph or less) written explanation of the results you obtained from Scenario 4. Did your estimated population size deviate from the truth once we violated the closure assumption, and if so why do you think that happened? What was the consequence of the assumption violations in terms of the model inputs and outputs? Please upload the answer to your question as a Word document. Also use this Word document to provide your population growth rate estimates for scenario 5.

Your completed Excel file, Word document, and R script will be due as uploads on Brightspace prior to the beginning of next week's lab. Importantly, in order to grade the assignment we need to be able to evaluate whether you correctly collected the data and accurately derived the estimates for $\hat{N}$ and its confidence intervals. So, please label your spreadsheet appropriately so that we can follow your work flow. As a reminder, all work should be shown by using formula within Excel to make calculations (i.e. don't do the math outside of excel and just enter the solutions), and your worksheet sheet should follow the guidelines listed above. We will grade your assignment based on the accuracy of your calculations, how closely your figures follow the formatting guidelines above, and whether or not your spreadsheet is easy to interpret.

8 Lab Appendix

8.1 Quick reference for important R commands in this lab

For this first lab we didn't use many specific R commands that will be terribly relevant to future weeks - the main goal was to give you a little more comfort in executing code and developing script in RStudio. But, we did learn that creating an object overwrites a previous object with the same name. Avoid this by using a logical naming convention if you create multiple versions of the same object

```
## Instead of this

Object.name <- 500

Object.name

Object.name <- 300

Object.name


## Do this

Object.name.1 <- 500

Object.name.1

Object.name.2 <- 300

Object.name.2
```

8.2 For more information

For a very nice overview of best practices when working in spreadsheets:

Broman, K.W., and K. H. Woo. 2017. Data organization in spreadsheets. The American Statistician. 72:2-10 DOI:10.1080/00031305.2017.1375989 Available at this link

8.3 Glossary of Terms

Asymptotic Bias - A situation where model estimates become biased towards larger values when sample sizes become small, normally because of a mathematical artifact in the model equation that is only affected by small values.

Batch Marking - A marking technique in which animals (or plants, for that matter) are given a mark that is not unique to the individual. Batch marks may be unique to cohorts - for example, use of different color marks among study years.

Closed Capture Model - Any mark-recapture model in which closure is assumed among sampling periods, such that all marked individuals are always available for detection during the study period.

Individual Marking - A marking technique in which each animal may be uniquely identified.

Open Capture Model - A model which accommodates loss of marked animals, for example due to mortality or emigration, among sampling intervals.