23 POPULATION ESTIMATION: MARK-RECAPTURE TECHNIQUES

Objectives

- Simulate the process of mark and recapture of individuals in a closed population.
- Estimate abundance using the Lincoln-Petersen method.
- Perform a Monte Carlo simulation to estimate the accuracy of the Lincoln-Petersen results.
- Determine how the number of individuals marked and number of individuals recaptured affects the precision of the Lincoln-Petersen index.
- Evaluate how emigration and capture probability can bias the Lincoln-Petersen index.

INTRODUCTION

How many moose are in Vermont? What is the population size of breeding black ducks in the Adirondacks? How many jaguars are in the Calakmul Biosphere Reserve in Mexico? How "confident" are we in our estimates? Estimating abundance in animals is a very common procedure for ecologists and land managers. This is because the size of a population can profoundly affect, among other things, its genetic make-up, probability of persistence, and rates of immigration, emigration, birth, and survival.

There are two basic ways of determining population size. The first is an actual "head count" of individuals, or **census**; the second is **estimation of population size through sampling**. The second method is the only option when (as is often the case) counting all individuals is impractical or impossible. There are different strategies for estimating plant and animal population sizes over time. The foremost difference is that animals move from location to location, whereas plants remain rooted in place and are thus often (but not always!) easier to count.

Because most animals are mobile, animal abundance is often estimated through mark-recapture techniques (Lancia et al. 1994). Deer, for example, are often marked with ear tags, and birds can be marked with color-coded bracelets attached to their legs. Marked animals are released and move freely about the population. A follow-up recapture session involves capturing a random sample of individuals from the population. Some individuals will contain markings, some will not. Mark-recapture techniques are based on the notion that the proportion of marked individuals in the second sample should be approximately equal to the proportion of

marked animals in the total population. In other words, if you know the number of marked and unmarked individuals captured in the second sampling session, and you know the number marked in the first sampling session, you can estimate the original population size in the first sampling session.

Several different mark-recapture models exist, including the Lincoln-Petersen model, the Schnabel model, and the Jolly-Seber model. Of these, the Lincoln-Petersen method is the simplest, involving only a single marking session and a single recapture session. This procedure was used by C. J. G. Petersen in studies of marine fishes and by F. C. Lincoln in studies of waterfowl populations (Seber 1982). The data in the model include the number of individuals marked in the first sample (M); the total number of individuals that are captured in the second sample (C); and the number of individuals in the second sample that have markings (R). These data are used to estimate the total population size, N, as

$$\frac{N}{M} \approx \frac{C}{R} \qquad \text{Equation 1}$$

Let's assume we are trying to estimate the population size of ladybug beetles in a given area. Equation 1 says that the ratio of the *total* number of ladybugs in the population to the total number of *marked* ladybugs is equal to the ratio of the number of ladybugs in the *sample* to the number of *marked (recaptured)* ladybugs in the sample. We can rearrange Equation 1 to get an estimate, of the total population size:

$$\hat{N} = \frac{CM}{R} \qquad \text{Equation 2}$$

This formula is the Lincoln-Petersen index of population size. In our spreadsheet, we will allow resampling (that is, an individual may be recaptured more than once). In this situation, the following modified index provides a better overall estimate of the population size when multiple trials are conducted:

$$\hat{N} = \frac{M(C+1)}{R+1} \qquad \text{Equation 3}$$

The Lincoln-Petersen estimate assumes that the population is **closed**—that immigration and emigration are negligible and the population does not change in size between the mark and recapture sessions. Other assumptions include:

- The second sample is a random sample.
- Marking does not affect the recapture of individuals.
- Marks are not lost, gained, or overlooked.

The Schnabel model is similar (in theory) to the Lincoln-Petersen method but involves more than one mark and recapture episode. The Jolly-Seber model relaxes the assumption that the population is closed (see Krebs 1999 for an overview of these methods).

Once we have an estimate of population size, it's critical to determine just how *confident* you are in your estimate. After all, you *will* arrive at an estimate, but since all sampling involves error, your estimate is probably off target by some amount. In this exercise, you will use a Monte Carlo simulation to get a feel for the range of values returned by the Lincoln-Petersen index. A **simulation** is any analytical method meant to imitate a real-life system. A **Monte Carlo simulation** is a statistical technique in which a quantity is calculated repeatedly, using randomly selected "what-if" scenarios for each calculation. In a nutshell, the technique uses a data-generating mechanism (such as the random number function in a spreadsheet) to model a process you wish to understand (such as the "behavior" of the Lincoln-Petersen index, when, for example, $M = 20$ and $C = 30$). New samples of simulated data are generated repeatedly, and the results approximate the full range of possible outcomes. The likelihood of each possible result can then be computed. The Monte Carlo technique derives its name from the casinos of Monte Carlo in Monaco, where the major attractions are games of chance and the successful gamblers must constantly calculate the probabilities of multiple possible scenarios in their heads.

PROCEDURES

In this exercise, you'll simulate a mark and recapture of individuals in a population of size 100 (the number you are trying to estimate). You'll calculate the Lincoln-Petersen index of abundance, run a Monte Carlo simulation to see the range of possible outcomes, and examine how the estimate and confidence intervals change as sample effort changes and as assumptions to the model are violated. Once you are an expert at Monte Carlo simulations, you can use the procedure to determine the best strategy for winning money at blackjack and head to Las Vegas (or better yet, Monaco).

As always, save your work frequently to disk.

INSTRUCTIONS

ANNOTATION

A. Set up and mark the model population.

For the sake of this exercise, we will consider a population of 100 individuals. However, you, the field biologist, don't *know* the actual population size is 100—you are trying to estimate it using the mark-recapture technique. You have been granted funding to mark 20 individuals. (We'll explore what happens if you mark fewer or more individuals later in the exercise.)

1. Open a new spreadsheet and set up column headings as shown in Figure 1.

	A	B	C	D	E
1	***Population Estimation - Lincoln-Petersen Mark-Recapture Model***				
2					
3					**Initial Sampling**
4		**M = number of new individuals marked =**			20
5				**Marking =**	m
6		**probability of remaining in population =**			1
7			**probability of recapture =**		1
8					
9	**MARK**		**RECAPTURE**		
10	**Individual #**	**Marking**	**Individual**	**Marking**	**C**

Figure 1

2. In cell E4, enter the number of individuals you will mark.

Enter the number 20 in cell E4.

3. Enter the letter **m** in cell E5.

The mark you will give to the 20 individuals is the letter **m**. The unmarked individuals will have the letter **u** associated with them.

4. Enter 1 in cells E6 and E7.

The Lincoln-Petersen method assumes that the population is closed (births, deaths, emigration, and immigration are negligible) and that all individuals have the same probability of capture and recapture. The values in cells E6 and E7 will allow us to explore violations of these assumptions. Cell E6 is the probability that an individual will remain in the population. For now it is set to 1 to meet the assumption that the population is closed. If individuals leave the population, either through death or emigration, that probability will decrease. Cell E7 is the probability that an individual will be recaptured, which we will also set to 1. If certain individuals (either marked or unmarked) tend to avoid traps in the recapture session, that probability will decrease. Perhaps they have learned trap locations and have become "trap shy."

5. Set up a linear series from 1 to 100 in cells A11–A110.

Enter 1 in cell A11.
Enter **=1+A11** in cell A12. Copy your formula down to row 110.
This assigns a number to each individual in the population.

6. In cells B11–B110, enter an **IF** formula to mark the first 20 individuals with an **m**, and designate the remainder **u** (unmarked).

Enter the formula **=IF(A11<=$E$4,$E$5,"u")** in cell B11. Copy this formula down to cell B110.
This formula tells the spreadsheet to examine the number in cell A11. If that number is less than or equal to (**<=**) to value in cell E4 (i.e., **20**), return the marking listed in cell E5 (i.e., **m**); otherwise, return the letter **u**.

7. Save your work.

B. Simulate the recapture of individuals.

Now we have a sample of marked individuals that have been released back into the population, and we can (after a period of time) resample the population and compute the Lincoln-Petersen index. First we will "reshuffle" the population, draw individuals from the population at random, and determine whether the individuals are marked or not.

1. In cell C11, generate a random number between 1 and 100. Copy your formula down to row 110.

Two different formulae can be used to generate a random number between 1 and 100:

- **=RANDBETWEEN(1,100)**
- **=ROUNDUP((RAND()*100),0)**

The **RANDBETWEEN** formula is fairly straightforward. If this function is not available in your spreadsheet package, the second formula will work by generating a random number between 0 and 1 (the **RAND()** portion of the formula), multiplying the number by 100 (***100**) and rounding the result up to **0** decimal places.

2. Use the **AND** function in cell G3.

Enter the formula **=AND(RAND()<=$E$6,RAND()<=$E$7)** in cell G3.
We'll take a moment to learn about the **AND** function, which we'll use as part of the formula in the next step. The **AND** function evaluates conditions you specify and returns the word "true" only if *all* the conditions you specify are true, and the word "false" if *any* of the conditions are not true. It has the syntax **AND(condition1, condition2, . . .)**. The formula in cell G3 generates two random numbers between 0 and 1 (the **RAND()** portion of the formula). The conditions are that the first random number must be less than or equal to the value in cell **$E$6** (the probability of remaining in the population), and that the second random number must be less than or equal to the value in cell **$E$7** (the probability of being captured in the second sampling bout). Since cells E6 and E7 are currently set to 1, both random numbers will be less than or equal to 1, so the program will return the word "true."

Now set cell E6 and E7 to 0.7 and press F9, the calculate key, to see how this formula works. Occasionally, a random number greater than 0.7 will be drawn, and the program will return the word "false." When you are satisfied that you understand how the **AND** function works, return cells E6 and E7 to the value 1 and continue to the next step.

3. In cells D11–D110, enter a formula to determine whether or not a recaptured individual was marked.

Enter the formula **=IF(AND(RAND()<$E$6,RAND()<$E$7), VLOOKUP($C$11:$C$110,$A$11:$B$110,2),".")** in cell D11. Copy this formula down to cell D110.
Now we are ready to determine if the individual that was sampled was marked or not. We also need to determine if the individual that was sampled left the population through death or emigration (cell **$E$6**) and if the individual is trap-shy (cell **$E$7**). The formula in cell D11 is a combination of four functions: **IF**, **AND**, **RAND**, and **VLOOKUP**. Keep in mind that Excel performs the innermost functions first and then moves to the outer functions.

The two inner functions are **RAND()** functions, which draw a random number between 0 and 1. The first random number is compared to the value in cell **$E$6**, which is the probability that the individual remains in the population. The second random number is compared to the probability that the individual is not trap-shy. In order for an individual to be captured in the recapture session, both probabilities need to be considered; this is done with the **AND** function, which will return "true" only if the individual stays in the population and is not trap-shy. Now we are ready for the **IF** function. If the individual remained in the population and was not trap-shy, then Excel moves to the **VLOOKUP** function. However, if the individual either left the population through death or emigration, or was trap-shy, Excel returns a missing value in the cell (**"."**).

The **VLOOKUP** formula searches for a value in the leftmost column of a table and then returns a value in the same row from a column you specify in the table. It has the syntax **VLOOKUP(lookup_value, table_array, col_index_num, range_lookup)**. So, assuming the individual was indeed captured, Excel will look up the value given in column C (the shuffled individuals) in a table given in columns A and B (specifically, cells A11–B110), and will return the value in the second column of the table (**m** or **u**); note that the **range_lookup** parameter is optional, and we are leaving it blank. In other words, assuming the individual is still in the population and can be captured, the **VLOOKUP** formula will find its number in column A and relay its marking from column B.

4. In cells E11–E110, sum two COUNTIF formulae to tally *C*, a running tally of the number of marked (**m**) plus unmarked (**u**) individuals recaptured.

Enter the formula **=COUNTIF($D$11:D11,"u")+COUNTIF($D$11:D11,"m")** in cell E11. Copy the formula down to cell E110.
To calculate *C* in column E, we count the individuals that are marked and those that are unmarked, then sum the two together. Remember to "anchor" the first reference to cell D11 with dollar signs (absolute reference). Also remember to use quotes around the letters u and m since they are nonnumerical data. Note that when cells E6 and E7 are both set to 1 (i.e., when the population is closed and no individuals learn to evade recapture), this formula will simply produce a linear series from 1 to 100 in column E. When the value in E6 or E7 is less than 1, however, not every capture attempt in column D will result in capturing an individual, so we will need this column to keep track of those that do.

5. Press F9, the calculate key, to simulate recapture outcomes.

C. Calculate and graph the Lincoln-Petersen index.

1. Set up column headings in cells F9–G10 as shown.

	F	G
9	**PETERSEN ESTIMATE**	
10	**R = total recaps (m)**	**Petersen Est**

Figure 2

2. In cells F11-F110, calculate *R*, the cumulative total number of recaptures.

To calculate the Lincoln-Petersen index, we need to keep track of *M*, *C*, and *R*. We'll assume that we start to recapture individuals one at a time, and we'll calculate the Lincoln-Petersen index each time a new individual is captured. The number marked, *M*, is given in cell E4. The numbers captured in the second session, *C*, are given in column E. A count of the number recaptured that were marked (*R*) will be tallied in column F. Row 11 simulates our first capture (Figure 3). We need to determine if the individual was marked or not, and then keep a running tally of recaptured individuals as we continue to capture individuals.
Enter the formula **=COUNTIF($D$11:D11,"m")** in cell F11. Copy your formula down to row 110.

	C	D	E	F
10	**Individual**	**Marking**	**C**	**R = total recaps**
11	12	m	1	1
12	95	u	2	1
13	31	u	3	1
14	2	m	4	2
15	2	m	5	3
16	88	u	6	3

Figure 3

3. Calculate the Petersen estimate in cells G11-G110.

Enter the formula **=($E$4*(E11+1))/(F11+1)** in cell G11. Copy the formula down to cell G110.
This is the spreadsheet version of Equation 3:

$$\hat{N} = \frac{M(C+1)}{R+1}$$

4. Graph the Lincoln-Petersen index as a function of *C*, the number of individuals captured in the second sampling bout.

Use the line graph option and label your axes fully. Your graph should resemble Figure 4.

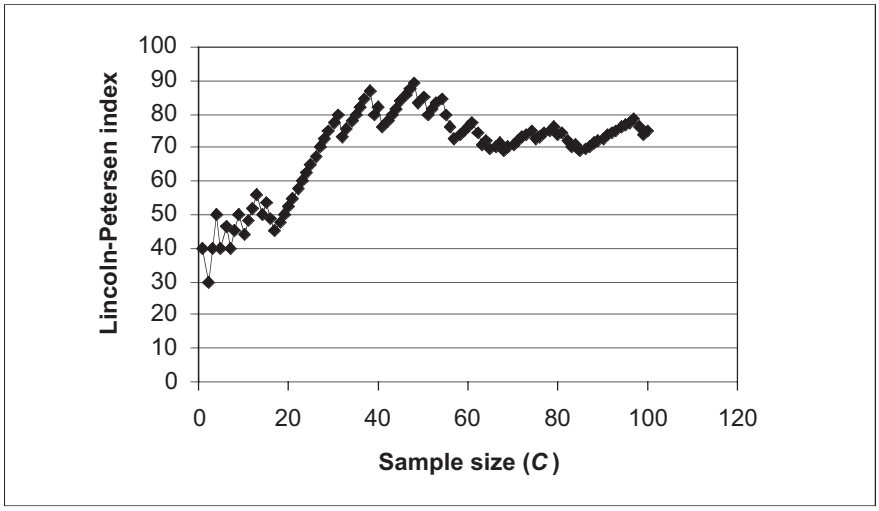

Figure 4

5. Answer questions 1–3 at the end of the exercise before proceeding.

D. Perform a Monte Carlo simulation.

Let's suppose that we mark 20 individuals and capture 20 individuals in the second sampling bout. How much confidence can we place in the resulting Lincoln-Petersen estimate? In this section we will set up a Monte Carlo simulation to see the range of estimates returned by our Lincoln-Petersen index. To do this, we will need to repeat our entire exercise 1000 times, each time generating a new index. Then we will examine how the index "behaves" based on our 1000 trials. We'll write a macro and let the computer do the tedious work for us.

1. Return the value in cell E4 to 20 individuals marked.

2. Set up new column headings in cells I9–O10 as shown in Figure 5.

	I	J	K	L	M	N	O
9	MONTE CARLO SIMULATION						
10	**Trial**	**L-P index**	**Low 2.5%**	**High 2.5%**	**Mean**	**Summary**	

Figure 5

3. Set up a linear series from 1 to 1000 in cells I11–I1010.

Enter 1 in cell I11.
Enter **=1+I11** in cell I12. Copy the formula down to cell I1010.

4. Set the calculation key to manual.

Open Tools | Options | Calculation and select Manual.

5. Develop a macro to run a Monte Carlo simulation.

Bring your spreadsheet macro program into record mode and assign a name and shortcut key (we used the shortcut <Control>+<m>).

Figure 6

If the small Stop Recording toolbar (Figure 6) doesn't automatically appear, open View | Toolbars | Stop Recording. The filled square on the left is the "stop recording" button, which you press when you complete your macro. The button to the right is the **relative reference button**. By default the button is "off," as shown above, which means that your macro records keystrokes as absolute references. Leave the button off *for now* and record the following steps:

- Select cell E10.
- Press F9, the calculate key, to generate new random numbers and hence a new simulation of mark-recapture.
- Open Edit | Find. Enter the number 20 in the box labeled Find What as shown in Figure 7. Select the Search by Columns and Look in Values options. Click the Find Next button, then Close. Excel will move your cursor down to the 20th individual captured.

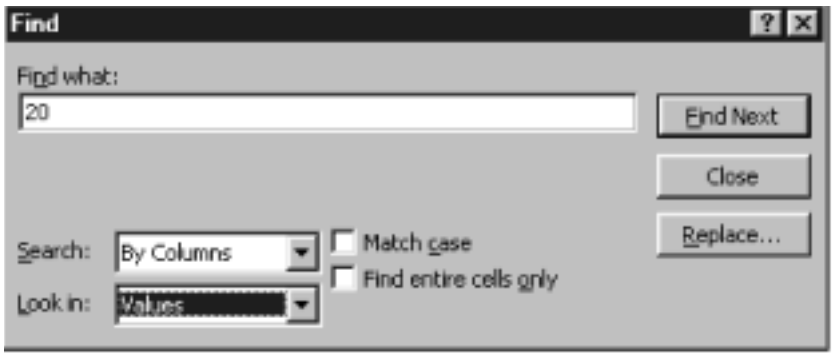

Figure 7

- Press the relative reference button (see Figure 6); it should become a lighter shade when depressed. Excel now assumes that cell references are relative rather than absolute.
- Use the right arrow key to move your cursor two cells to the right. This cell holds the Lincoln-Petersen estimate associated with 20 captured individuals in the second session and a variable number of marked and recaptured individuals.
- Click the relative reference button off.
- Open Edit | Copy.
- Select cell J10.

- Open Edit | Find. Leave the Find What box blank and Search by Columns. Click the Find Next button, then Close.
- Open Edit | Paste Special. Then select the Paste Values option. Press OK.
- Click on the Stop Recording button.

Now when you press your shortcut key 1000 times you will generate 1000 new Lincoln-Petersen indices, each one generated by random numbers and following the parameters established in the model. A simple shortcut outlined in the next step can save you 1000 keystrokes.

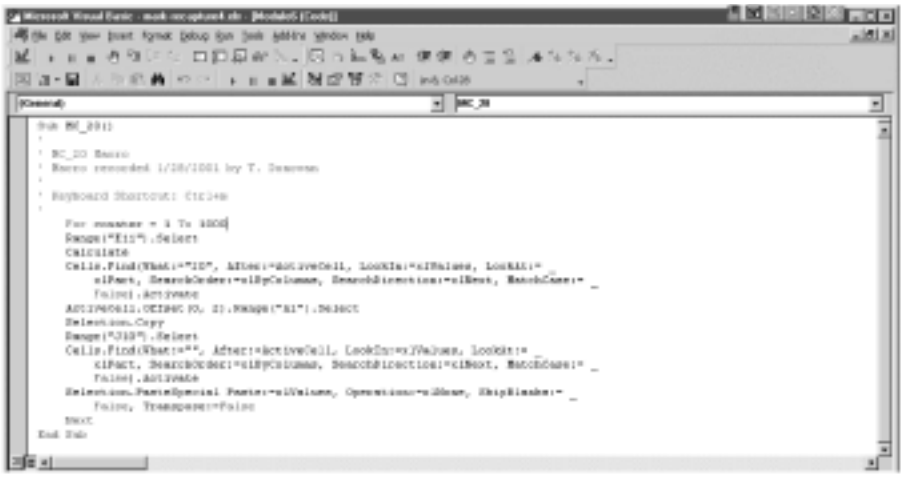

Figure 7

6. (Optional) Edit your macro using the Visual Basic code.

To edit your macro, open Tools | Macro | Macros, and click the Edit button. You should see a box that reveals the Visual Basic Applications code that Excel recorded as you entered your macro (Figure 7)

- After the `Keyboard Shortcut Control+m`, press Return and type in the words **`For counter = 1 to 1000`**
- Before the last line of code, which reads `End Sub`, create a new line and type in the word **`Next`**. Close out of the box to return to your spreadsheet.

Now you press <control>+<m> just once and your new macro, which consists of 1000 different simulations, will run. Before running the macro, you should delete any previous results from column J (otherwise you will wind up with more than 1000 results in this column). You can do this by highlighting any results in this column and pressing the Delete key.

When you press <control>+m, your computer will flash for several minutes as it cranks through the simulation. Caution: If you use another program while the simulation is running, be careful not to copy material to the clipboard—the simulation is making extensive use of the clipboard (through copy and paste), so putting other material there can cause errors.

7. Examine your results from 1000 trials.

Bear in mind that in actual mark-recapture experiments we don't know the total population size—that's what we're trying to estimate. This Monte Carlo simulation allows us to determine, for the special case in which $N = 100$, just how likely the Lincoln-Petersen index is to come up with an "acceptable" estimate. What is acceptable will depend on the purpose of the experiment (see question 4 at the end of the exercise).

When analyzing results, scientists like to be at least 95% certain that a given result is not due to chance. You can use your spreadsheet to see the range of values that the Lin-

coln-Petersen estimate will return 95% of the time. Since you have 1000 results from your Monte Carlo simulation, the "middle" 950 values represent this range. The remaining 50 values are the 25 highest and 25 lowest Lincoln-Petersen estimates from your simulation. We are interested in determining the 25th highest observation and the 975th highest observation. The **LARGE** function does this: it returns the *k*th largest value in a data set—you specify the data set and what value you want returned.

8. In cell K11, compute the value of the 975th highest estimate.

Enter the formula **=LARGE(J11:J1010,975)** in cell K11.

9. In cell L11, compute the value of the 25th highest estimate.

Enter the formula **=LARGE(J11:J1010,25)** in cell L11.

10. In cell M11, compute the average Lincoln-Petersen index from your simulation.

Enter the formula **=AVERAGE(J11:J1010)** in cell M11.

E. Optional: Generate descriptive statistics on your results.

This step requires that that the Analysis ToolPak be activated. To activate the ToolPak, go to Tools | Add-Ins and click on the ToolPak option, then press OK. To generate descriptive statistics, go to Tools | Data Analysis | Descriptive Statistics. The dialog box in Figure 8 will appear.

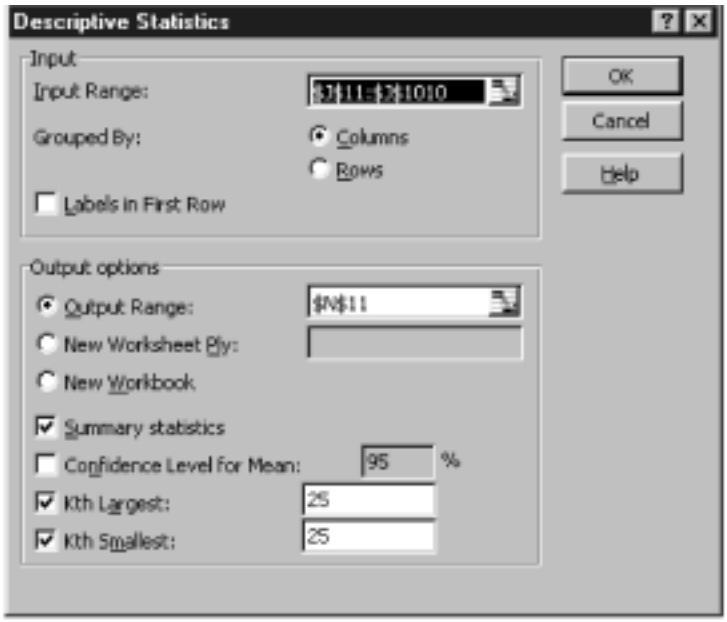

Figure 8

The Input Range will be the results of your 1000 simulations (you can use the relative reference button to enter this). Use **$N$11** as the output range, check the Summary statistics option, and enter **25** as the Kth Largest and Kth Smallest values. Excel will return descriptive statistics in columns N and O, as shown in Figure 9. The 95% confidence intervals are obtained by examining the 25 highest and lowest values. The confidence values should match the values you computed in cells K11 and L11. Our simulation revealed that, for a population where 20 individuals are marked in the first session

and 20 individuals are captured in the second session, the Lincoln-Petersen index fell between 46.7 and 210 individuals 95% of the time. Your answer might be slightly different. Remember that the true population size is 100 individuals. You may be able to get better estimates by changing M and/or C.

	N	O
11	*Column1*	
12		
13	Mean	100.3880152
14	Standard Error	1.746782658
15	Median	84
16	Mode	84
17	Standard Deviation	55.23811776
18	Sample Variance	3051.249654
19	Kurtosis	13.9588804
20	Skewness	3.130416209
21	Range	385
22	Minimum	35
23	Maximum	420
24	Sum	100388.0152
25	Count	1000
26	Largest(25)	210
27	Smallest(25)	46.66666667

Figure 9

QUESTIONS

1. Based on your initial setting of $M = 20$, how does C (the number captured in the second sampling bout) affect the Lincoln-Petersen index? Press F9, the calculate key, to run several simulations and get a qualitative feel for the relationship.
2. Change the value in cell E4 to 50, then 70, then 90 to increase the proportion of the population that is initially marked. For each value, press F9 several times to get a general feel for the results. How does this increase in proportion of marked individuals affect the Lincoln-Petersen estimate? What happens to the Lincoln-Petersen estimate when 100 individuals are marked? Use graphs to illustrate your answer.
3. Examine your graph from Part E (the Lincoln-Petersen index as a function of C). How were the data collected to generate such a relationship? Is this a legitimate way to evaluate how the Lincoln-Petersen index changes as C increases? Why or why not?
4. Suppose you are planning to study population fluctuations of a species of frog living in a particular pond, and your initial "guesstimate" is that the pond currently has about 100 frogs living in it. Discuss the value of estimating variations in the population size by marking 20 individuals and recapturing 20 individuals. Try different values for M and C to try to determine an experimental design that will produce an "acceptable" margin of error. Which has a greater effect on the range of results: increasing M or increasing C?

   To change M, simple change the value in cell E4. To change C, you need to edit the macro: Open Tools | Macro | Macros, highlight your macro on the list that appears, and click on the edit button. In the macro editing window that opens,

the first "Find What" value represents *C*, and you can change it to any value up to 100. Remember to clear your results from column J before running the macro each time (or, if you want to keep your previous results, save your spreadsheet with a different name).

5. Set cell E4 equal to 50. How do violations of the assumptions of "closed" population and equal catchability affect the Lincoln-Petersen estimate? Set cells E6 to 0.6 (thus, 40% of the individuals leave the population) and set cell E7 to 0.7 (thus, 30% of the individuals are unlikely to be captured in the second sampling bout for some reason). Assuming that you can recapture 30 individuals ($C = 30$), how do the results of the Monte Carlo simulation change as a result of these violations?

*6. Assume that the population is closed (cell E6 = 1). Assume further that the probability of recapture pertains only to those individuals that were marked in the initial sampling period (perhaps the individuals have learned to avoid traps after being captured earlier). How could the model be modified to reflect this situation?

LITERATURE CITED

Krebs, C. J. 1999. *Ecological Methodology*, 2nd Ed. Benjamin/Cummings. Menlo Park, CA.

Lancia, R. A., J. D. Nichols and K. H. Pollock. 1994. Estimating the number of animals in wildlife populations. *In* T. Bookhout (ed.), *Research and Management Techniques for Wildlife and Habitats*, 5th Ed., pp. 215–253. The Wildlife Society, Bethesda, MD.

Seber, G. A. F. 1982. *The Estimation of Animal Abundance and Related Parameters*, 2nd Ed. Macmillian, New York.