6

SAMPLING SPECIES RICHNESS

Objectives

- Simulate a population of 1000 individuals composed of various species.
- Calculate species richness by sampling.
- Determine how community composition affects species richness estimates.
- Develop a bootstrap analysis of how sample size affects species richness estimates.

INTRODUCTION

Imagine you are a conservation biologist conducting surveys of insect species in previously unstudied areas. Your mission is to estimate the number of species occurring in different habitat types across a large region. The number of species that occurs in a particular area is called its **species richness**, and it is just one of many measures of biodiversity. A practice known as a *rapid biodiversity assessment* is currently being used by many conservation organizations to survey the biodiversity of plants and animals before pristine habitats are altered and developed (see, for example, http://www.conservation.org/RAP/Default.htm). Assume there are 10 locations that must be sampled in a short period of time. How many samples should you take at each site to estimate the number of insect species in a location before moving onto the next location? Time and funding are short and you will not be able to do a complete survey of the insect biota.

A basic problem is that it is nearly impossible to count every single species in a community. If funding and time were unlimited, you might conduct a complete census and enumerate all of the species in the community. However, this is not often the case; instead you must settle for *sampling* the community and estimating its species richness based on this sample of individuals. Estimating species richness by sampling presents some major challenges. First, you are likely to miss some species. And second, although the more you sample in a particular area the more likely you are to find new, previously unsampled species, there is a point of diminishing returns that must be considered in your sampling efforts.

For example, consider a community that consists of 1000 insect species, and you sample insects by sweeping the vegetation with a net. In your first sweep, you capture 25 species. In your second sweep, you capture 30 species, but 20 of

these were already captured in the first sweep. Thus, with 2 samples your total species richness is 35 (25 new species recorded with the first sweep, and 10 new species recorded with the second sweep). With each sweep (sample), the chances of adding a new, previously unsampled species decreases. At some point it becomes cost-effective to move
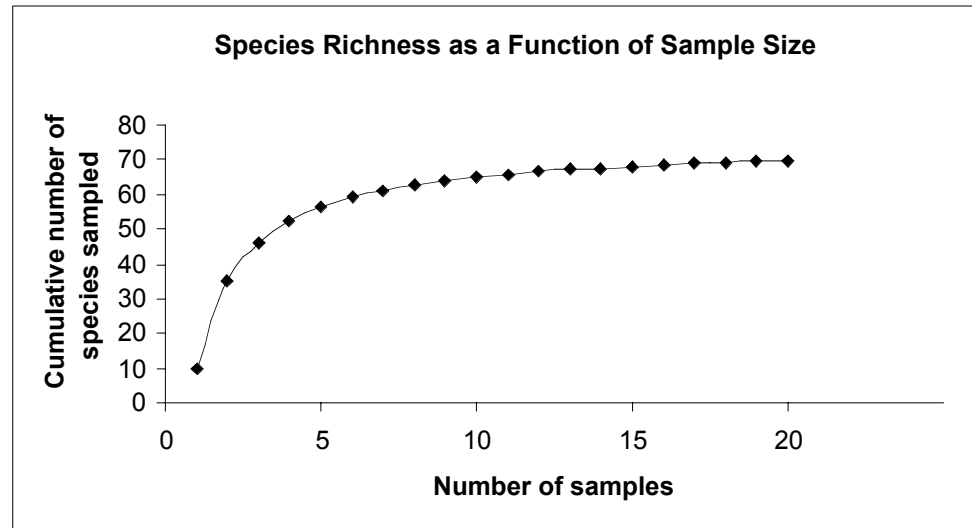

Species Richness as a Function of Sample Size

Figure 1

to the next location and start sampling anew. In the example shown in Figure 1, taking 15 samples will yield more or less the same species richness estimate as taking 18 or 20 samples.

What factors will determine the shape of a sampling curve such as Figure 1? One factor is the distribution of the individuals within the community. If the community consists of 100 species, but 90% of the total individuals are from species 1, most of our samples will consist of species 1, and we may have to take many samples to encounter one of the rarer species. In contrast, if the numbers of individuals in the community are more or less evenly distributed across 100 species, so that no single species dominates the community, you may not have to sample as much because all species are equally abundant.

Another general problem with sampling is that you will never *really* know how well your species richness estimate measured the *true* species richness in a community. After all, this is what you are trying to estimate with your sampling. With advances in computing, however, it is now possible to ask the question, *"If we take a different, random sample from a community with a known number of species, how does the species richness estimate change as sample size changes?"* The difference between the actual species richness of the community and the estimated species richness based on sampling is called **bias**.

One method for analyzing bias is a **bootstrap analysis**, which involves taking random samples of the data (with replacement so that the same individuals can be sampled more than once), calculating the parameter of interest (in this case, species richness), repeating the process for 1,000 or more trials for a given sample size, and then estimating the mean and standard deviation of species richness from the replicate bootstrap estimates. As discussed in Exercise 4, this process is relatively straightforward with spreadsheets.

Since the number of species in the community in your bootstrap analysis is known *a priori* (known beforehand), the bootstrap analysis gives you an indication of how sample size, as well as community composition, biases your estimate of species richness. The purpose of this exercise is to introduce you to sampling and bootstrap methods as they pertain to species richness. As always, save your work frequently to disk.

INSTRUCTIONS

ANNOTATION

A. Set up the model community.

1. Open a new spread-sheet and set up column headings as shown in Figure 2.

	A	B	C	D	E	F
1	*Sampling Species Richness*					
2						
3			**Tally**			
4	**Species**	**# in pop**	0			
5	1	100				
6	2	100				
7	3	100				
8	4	100				
9	5	100				
10	6	100				
11	7	100				
12	8	100				
13	9	100				
14	10	100		**<-- This number must equal 1000.**		
15						
16	**Total =**	1000				

Figure 2

We will consider a community in which there are 1000 total individuals and up to 10 different species. The species identification is given in cells A5–A14. The numbers of individuals of each species are given in cells B5–B14.

2. Enter the values shown in cells B5–B14.

To begin, let's consider a community that is evenly distributed with 100 individuals of each species. Later in the exercise, you will be able to change the composition of the community by altering the values in cells B5–B14.

3. In cell B16, enter a for-mula to sum the total number of individuals in the community.

Enter the equation **=SUM(B5:B14)** in cell B16. Your result should be 1000.

4. Graph the distribution of the 1000 individuals among the 10 species. Use a column graph, and label your axes fully (Figure 3).
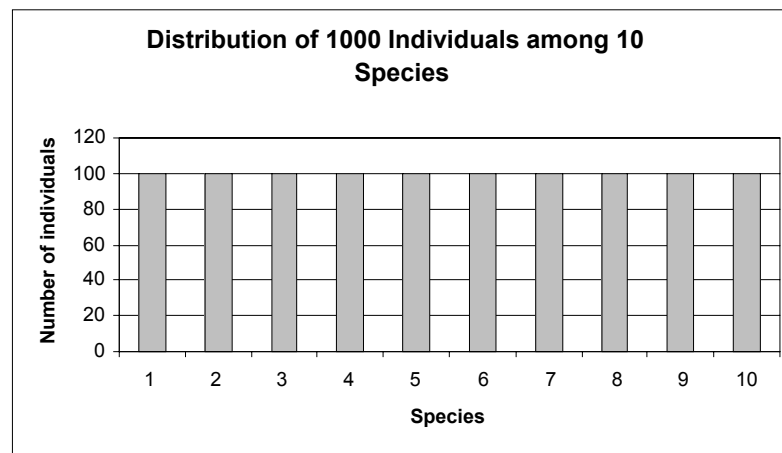
Figure 3

5. Compute a "running tally" of individuals in C4–C14.

Enter 0 in cell C4.
Enter the formula **=B5+C4** in cell C5 and copy this formula down to cell C14.
The formula in cell C5 gives the tally of individuals when only the first species, species 1, has been considered. Copying the formula down the column keeps a running tally of the number of individuals in the community as more species are observed. The result in cell C14 should be 1000, to account for all of the individuals present in the community. This "tally" will allow you to assign a species identification to individuals in a later step.

6. Save your work.

B. Sample from the community and compute species richness.

Now we are ready to sample from this community (one individual at a time) and estimate species richness. Since there are 10 species present (each with 100 individuals), species richness is 10. You will try to estimate this parameter by randomly sampling the population and computing richness.

1. Set up new spreadsheet headings as shown in Figure 4.

	A	B	C	D	E	F
26				**Random sample**		
27	**Individual**	**Species**	**Sample size**	**Individual**	**Species**	**Richness**

Figure 4

2. Set up a linear series from 0 to 999 in cells A28–A1027.

Enter 0 in cell A28.
Enter **=1+A28** in cell A29. Copy this formula down to cell A1027.
This series will represent the 1000 individuals in the community.

3. In cell B28, use the **LOOKUP** function to assign a species to the individual in cell A28. Copy this formula down to cell B1027.

Now we will identify which species each individual belongs to, based on the species identification (1–10) given in column A and the tally given in cells C4–C14. In cell B28, enter the formula **=LOOKUP(A28,$C$4:$C$14,$A$5:$A$14)**. The **LOOKUP** function looks up a value (the value in cell A28) in a vector that you specify (**$C$4:$C$14**), and returns a value from a corresponding vector (**$A$5:$A$14**). (A *vector* is a single row or column of values). In this case, it compares the value in cell A28 (which is 1) to the values in cells C4–C14; it finds that A28 is equal to 0 (the value in **$C$4**), so it returns the value in **$A$5**, which is 1. In other words, it assigns individual 1 to species 1. (Note that with this formula, the value in the tally and the species assignments are offset by one row.)

The **LOOKUP** function is handy for assigning species to individuals because if the function can't find the exact lookup value, it matches the largest value in the lookup vector (cells C4–C14) that is less than or equal to the lookup value. For example, when it looks for individual 449 in **$C$4:$C$14**, the largest value it can find that is less than 449 is 400, so it will assign this individual to species 5 (the value in **$A$9**, which is the cell corresponding to **$C$8**).

The result is that species are assigned to individuals with the distribution you determined in cells B5–B14. Your first 100 individuals should all be species 1, the next 100 individuals should all be species 2, and so forth. To test the function, set cell B6 to 1000 and set the remaining cells in B5–B14 to 0. Remember that the final tally of individuals must equal 1000 in cell C14. All 1000 individuals should now be species 2. When you feel you have a handle on how the **LOOKUP** function works, return cells B5–B14 to 100, and continue to the next step.

4. Set up a linear series from 1 to 1000 in cells C28–C1027.

Enter 1 in cell C28.
Enter **=1+C28** in cell C29. Copy this formula down to cell C1027.

5. In cells D28–D1027, generate a random number between 0 and 999 to designate a randomly sampled individual in the population.

Enter the formula **=ROUND(RAND()*1000,0)** in cell D28. Copy this formula down to cell D1027.

Cell D28 represents the first individual sampled, cell D29 represents the second individual sampled, and so on. Note that an individual can be sampled more than once if the same random number is drawn.

The **RAND()** function generates a random number between 0 and 1. When the random number is multiplied by 1000 and then rounded to 0 decimal places with the **ROUND** function, the result is a randomly sampled individual from the population. (If your program has the **RANDBETWEEN** function, the formula **=RANDBETWEEN(1,1000)** will do the same thing.)

6. In cell E28, enter a **LOOKUP** formula to identify the species of the randomly chosen individual in cell D28. Copy this formula down to cell E1028.

Enter the formula **=LOOKUP(D28,$A$28:$A$1027,$B$28:$B$1027)** in cell E28. Copy it down to cell E1028. Column E returns the species of each randomly selected individual. It uses another **LOOKUP** function to do this. The formula in cell E28 tells Excel to lookup the value in cell D28 (the randomly selected individual) in the vector of cells A28–A1027 and return this individual's species identification, given in cells B28–B1027.

7. Enter the number 1 in cell F28.

Finally we are ready to compute species richness—the total number of species—as our sampling progresses. Cell F28 is the first sample, so species richness will be equal to 1.

8. In cell F29, enter a nested **IF(COUNTIF())** formula to calculate the species richness, and copy this formula down to cell F1028.

With our second sample, we need to evaluate whether species richness is 1 (i.e., we sampled the same species in sample 2 as we did in sample 1) or 2 (i.e., we sampled a new species in sample 2). Enter the formula **=IF(COUNTIF($E$28:E28,E29)>0, F28,F28+1)** in cell F29. This is an **IF** formula with a **COUNTIF** formula nested within it. An **IF** formula has 3 parts to it, each separated by a comma. The first part is called the *criterion*. In this case, our criterion is **COUNTIF($E$28:E28,E29)>0.** The **COUNTIF** formula counts the number of times a certain value appears in a range of cells. Our formula tells the spreadsheet to examine cell E29 and count the number of times this value appears in the range of cells E28–E28. If this number is greater than 0 (the second sample was also recorded in the first sample), the program carries out the second part of the **IF** statement; if this number is not greater than 0, it carries out the third part of the **IF** statement. Thus, our example will look at the second species sampled (cell E29), and if this species number has appeared in the previous samples (E28–E28), the species richness value will remain at the previous number (cell F28); otherwise the richness will be increased by 1 (cell F28+1).
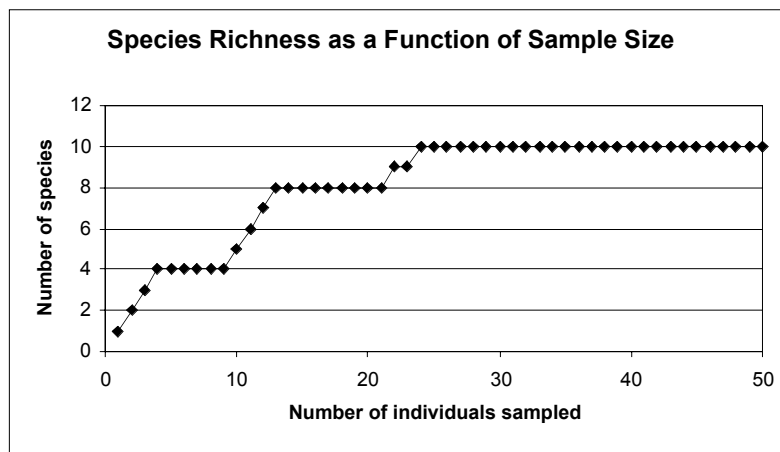
9. Graph species richness as a function of sample size. Use the scatter graph option, and label your axes fully (Figure 5).

Figure 5

Your graph will look different than ours because your random samples likely differed than ours. Keep in mind that the actual species richness of the community is 10 species. In our example, 24 individuals needed to be sampled to arrive at this number.

10. Press F9, the calculate key, a number of times to generate new samples.

Pressing F9 will generate new random numbers, and hence a new set of individuals that are sampled. With each simulation, you will notice that your species richness estimates change as samples accumulate. For example, a new simulation required over 40 individuals to be sampled to generate an unbiased estimate of species richness (Figure 6).
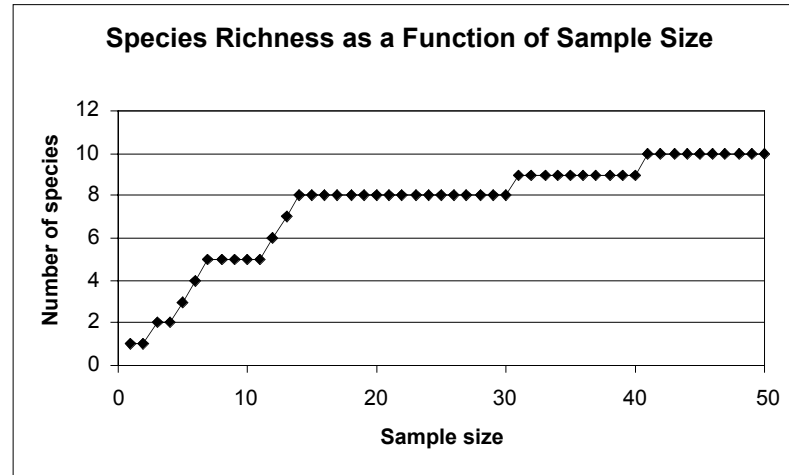

Species Richness as a Function of Sample Size

11. Save your work.

Figure 6

C. Set up the bootstrap.

The fact that each sampling simulation generates new and different results suggests the need for a bootstrap analysis. For example, if we took only 20 samples, how would our species richness estimate change from simulation to simulation? By "bootstrapping"—conducting many "replicate" sampling simulations—we can characterize the nature (mean and standard deviation) of our sampling with respect to species richness. We will do this for two of sample sizes ($n = 20$ and $n = 50$). We will run 1000 trials for each sample size, recording our species richness estimate with each simulation. This will provide useful information for deciding how many samples would be adequate at each location you need to sample.

1. Set up new column headings as shown in Figure 7.

	G	H	I	J	K
4		**Community 1**		**Community 2**	
5	**Trial**	$n = 20$	$n = 50$	$n = 20$	$n = 50$

Figure 7

2. Set up a linear series from 1 to 1000 in cells G6–G1005.

Enter 1 in cell G6.
Enter **=1+G6** in cell G7. Copy this formula down to cell G1005.

3. Create a macro to record species richness for sample size of 20 for 1000 trials.

First go to Tools | Options | Calculation and set your calculation key to Manual. Then put your Macro function in the "Record Macro" mode and assign a name and shortcut key. This macro provides one way to keep track of the species richness estimates when the sample size consists of 20 individuals. These estimates will be output into cells H6–H1005.

Record the following steps:
- Press F9, the calculate key, to generate a new set of random numbers, and hence a new set of randomly selected individuals.
- Select cell F47, the species richness estimate associated with a sample size of 20.
- Select Edit | Copy.
- Select cell H5, and then go to Edit | Find (Figure 8). Leave the Find What box completely blank; choose By Columns in the Search box and Values in the Look In box. Click Find Next and Close. Your cursor should move down to the next blank cell (trial 1).
- Go to Edit | Paste Special, and paste in Values, which is the species richness estimate for that trial.
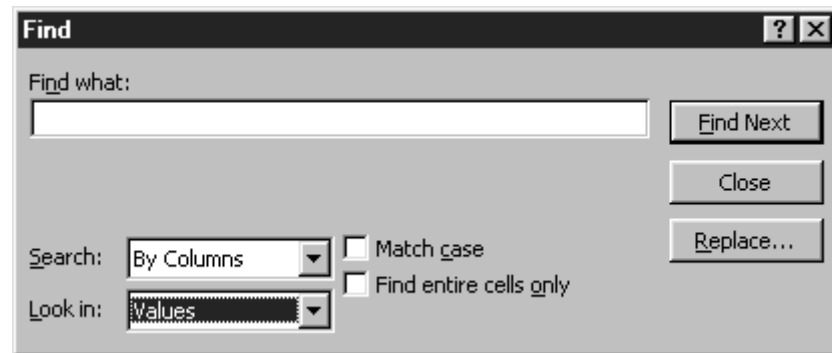- Select Tools | Macro | Stop Recording.

Figure 8

Now when you press your shortcut key, the macro will automatically conduct a new replicate sample and record the species richness values in the appropriate place. Run the macro 1000 times to complete your bootstrap analysis. This may take a while. If you like shortcuts, you can edit your macro's Visual Basic code by inserting two lines of code in the Visual Basic program, as follows:

- Open Tools | Macro | Macros.
- Click the Edit button to edit your macro called Trials. You should now see the Visual Basic code (Figure 9).
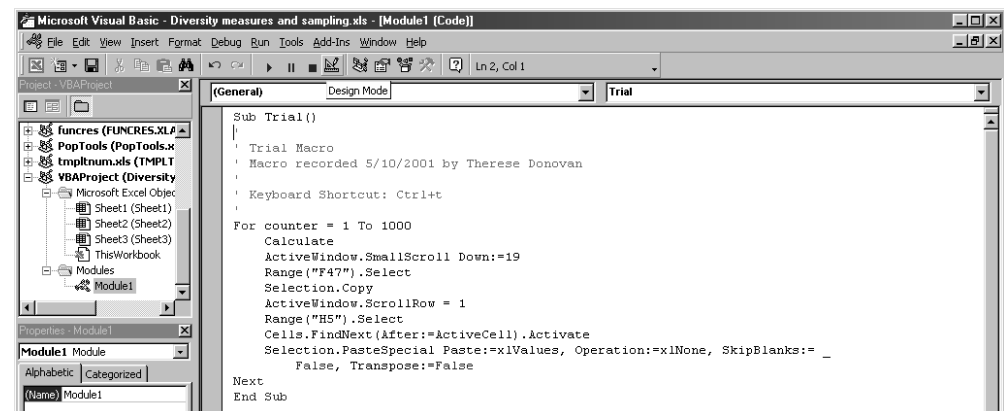

Figure 9

- Below line 4 (**Keyboard Shortcut**), enter a new line and type in the words **For counter = 1 to 1000** as shown in Figure 10.
- Above the last line (**End Sub**), enter a new line and type in the word **Next**.
- Exit the Visual Basic editor by clicking the close box in the upper right hand corner of the spreadsheet. You will be returned to your spreadsheet. Now when you press <Control>t, Excel will run 1000 trials for you.

4. Conduct a bootstrap analysis for a sample size of 50, and record the results of each bootstrap trial in column I.

You can record brand new macros, or edit the Visual Basic code in your existing macro. For the sample size of 50, you would highlight cell F77 (which is the species richness for a sample size of 50), and select cell I5 to record the results in the appropriate column. These slight adjustments can be made in the existing visual basic code. After you are finished, switch back to Automatic Calculation.

5. In cells H1006 and I1006, enter a formula to compute the mean species richness from the 1000 trials.

Enter the formulae
- H1006 **=AVERAGE(H6:H1005)**
- I1006 **=AVERAGE(I6:I1005)**

6. In cells H1007 and I1007, enter a formula to compute the standard deviation of species richness from the 100 trials.

Enter the formulae
- H1007 **=STDEV(H6:H1005)**
- I1007 **=STDEV(I6:I1005)**

7. In cells H1008 and I1008, enter a formula to divide the standard deviations by 2.

This step is necessary for graphing the standard deviations in the next step. Enter the formulae
- H1008 **=H1007/2**
- I1008 **=I1007/2**

8. Graph the mean species richness for the 1000 trials. Use a column graph and label your axes fully. Your graph should resemble Figure 10.

Mean Species Richness from 1000 Bootstrap Samples

Figure 10

9. Add the standard deviation bars to your graph.
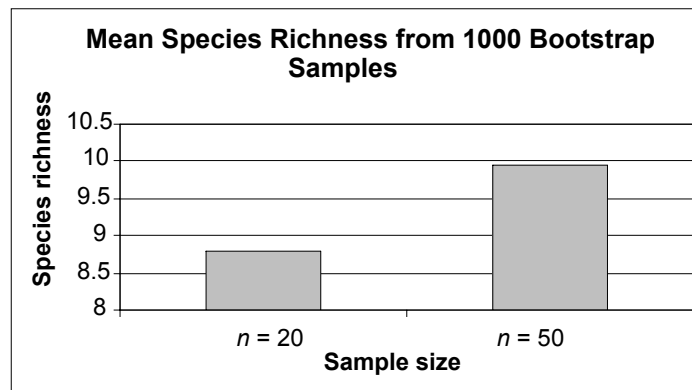
To add error bars, select the bars on the chart by clicking once on one of the bars. Then go to Format | Selected Data Series. A dialog box will appear (Figure 11).
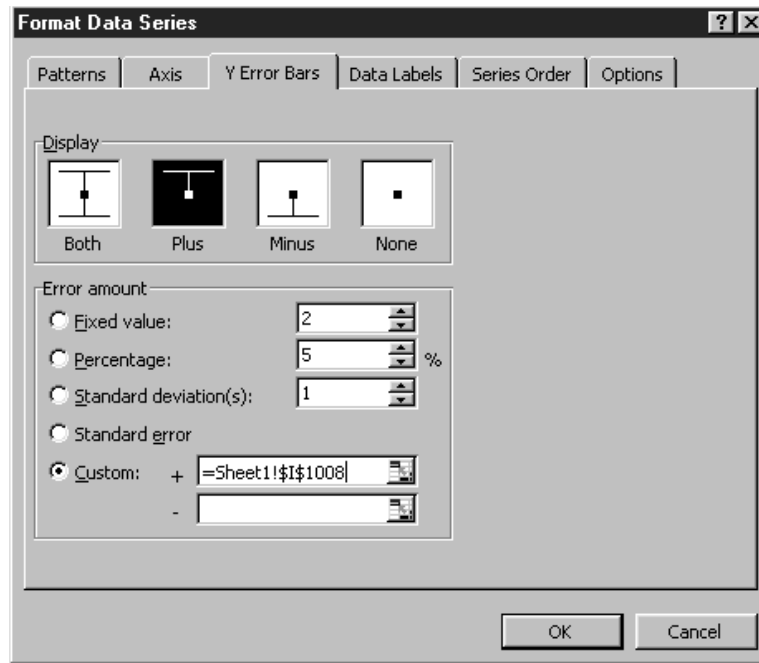
Figure 11

If you want to show only the top half of the errors, click on the Plus display, and then choose the Custom button. Then, in the window to the right of the + symbol, click on the little red arrow to shrink the box, use your mouse to select cell H1008, type in a comma, and use your mouse to highlight select cell I1008. Click again on the red arrow to bring the dialog box up again. Press OK and your graph should be updated (Figure 12). You should notice instantly that the larger sample size has a much smaller standard deviation than the smaller sample size, and that the larger sample provides a less biased estimate of species richness than the smaller sample. You must now consider the trade-offs between sampling a site intensively (n = 50 or more) at the expense of sampling a large number of sites.
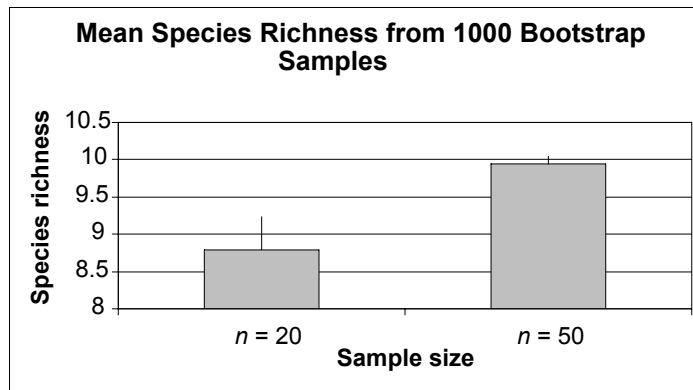

10. Save your work.   **Figure 12**

QUESTIONS

1. Fully interpret the last graph you created, the results of the bootstrap analysis for sample sizes of 20 and 50. Based on your results, is it worth sampling 50 individuals to ensure that your species richness estimate is unbiased?

2. How does the composition of the community affect species richness estimates? Set up your spreadsheet as follows:

	A	B	C
1	*Sampling Species Richness*		
2			
3			**Tally**
4	**Species**	**# in pop**	0
5	1	900	900
6	2	20	920
7	3	10	930
8	4	10	940
9	5	10	950
10	6	10	960
11	7	10	970
12	8	10	980
13	9	10	990
14	10	10	1000
15			
16	**Total =**	1000	

The new frequency distribution for species in this community should look like Figure 13. Develop a new macro, and sample from this new community with sample sizes of 20 and 50. Record your output under community 2 (columns J and K), and compare the bootstrap analysis for community 1 and community 2. Use graphs to explain your answer.
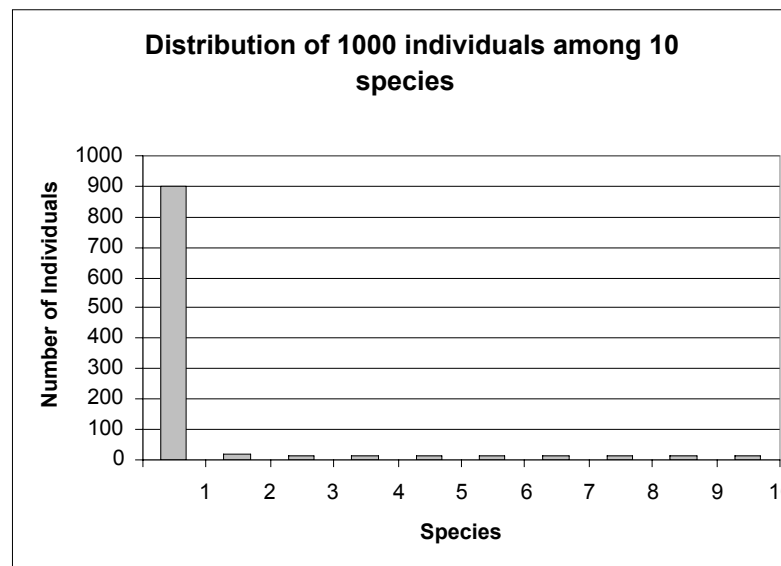

Distribution of 1000 individuals among 10 species

Figure 13

3. Species richness is only one measure of biodiversity for a community, but it is frequently used. Can you think of any shortcomings or assumptions of assigning conservation priorities to various locations based on species richness estimates?

LITERATURE CITED AND ADDITIONAL READINGS

Krebs, C. 1999. *Ecological Methodology*. 2nd Ed. Addison-Wesley Educational Publishers, Inc. Menlo Park, CA.

Moguel, P. and V. M. Toledo. 1998. Biodiversity conservation in traditional coffee systems of Mexico. *Conservation Biology* 13: 11–21.

Soberon, M. and J. B. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. *Conservation Biology* 7: 480–488.