

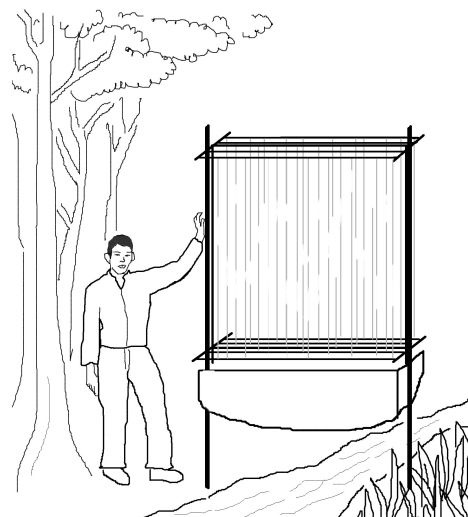
Estimating species richness with EstimateS

In this practical exercise we'll use the EstimateS software package to try to estimate the number of species of bats in a peat swamp forest in Malaysia.

A. Bat data from Loagan Bunut

The example data we'll use for this exercise are for bats caught in harp traps in peat swamp forest in Loagan Bunut National Park in Sarawak, Malaysia.

A harp trap has an aluminum frame with four parallel sets of vertical fishing lines (see the drawing on the left). A bat that flies into a harp trap will get caught in the fishing lines and slide down, uninjured, to the bag underneath. The bats are identified, weighed, measured, and released at the place they were captured at the end of the trapping session. Harp traps are very efficient in capturing echo-locating bats, typically forest interior species, which are unable to detect the thin lines of the harp trap.



The data are in the file "LBunut_bats_input.txt".

Open the file "LBunut_bats_input.txt" from in MS Excel® (or other spreadsheet program such as Calc from OpenOffice, <http://www.openoffice.org/>): either right-click on the file name and select "Open With > MS Excel" or drag-and-drop the file into an open MS Excel window. (If you use File > Open in MS Excel, the import wizard will start, which you don't need.)

The data file has a description of the data on the first line. On the second line is the total number of species (22) followed by the number of samples (50). Then come the records of bats captured, with a row for each species and a column for each sampling occasion, which shows the bats caught on one night at one location. For example, the first column shows that on the first night 6 species were caught (the other 16 rows contain "0"), with one species represented by two bats and the other five species represented by one bat each. The second sample had just two bats, both from the same species.

"LBunut_bats_input.txt" is formatted for input into EstimateS. Note that EstimateS does not use species names or sample identifiers.

B. Getting started with EstimateS

Go to the EstimateS web site (<http://purl.oclc.org/EstimateS>), and fill in the registration form. Then download the installer for the latest version of the software, currently EstimateS 8.0.0 in 'SetupEstimateSWin800.exe'. As with any .exe file, it's wise to save it on your hard disk, run a virus check with an up-to-date virus scanner, and create a Windows Restore Point before running the setup program. Run SetupEstimateSWin800.exe and follow the on-screen instructions.

If you use the results from EstimateS in any report or paper, please cite it as:

Colwell, R. K. 2006. EstimateS: Statistical estimation of species richness and shared species from samples. Version 8.0.0. User's Guide and application published at: <http://purl.oclc.org/estimates>.

Open EstimateS and click 'OK' to agree not to distribute EstimateS commercially. In EstimateS, click on 'File > Load Data Input File' (or press Ctrl-I). In the dialogue box, navigate to the file "LBunut_bats_input.txt" and click Open.

A confirmation screen appears with the title of the data set and the number of species (22) and samples (50). There's also a list of 'Optional' items: don't worry about these, as we can set options later.

Then a dialogue appears asking for the format of the input file. "LBunut_bats_input.txt" is Format 1 and there are no rows with sample information or columns with species names.

You should then see a box telling you that the data have been loaded successfully.

Go to 'Diversity > Diversity Settings...' (or press Ctrl-T) and then to the 'Estimators' tab. In the top section, click on the radio-button next to "Use classic formula for Chao 1 & Chao 2".

Then go to the 'Other Options' tab and check the box next to "Compute Fisher's alpha, Shannon, & Simpson indexes".

Leave the other settings as they are and click on 'Compute'.

If a box appears asking if it's OK to erase some old Diversity Statistics, click on 'OK'.

A box pops up briefly with a progress bar, then a table of results appears. EstimateS produces a huge mass of figures, far too much to digest. They make much more sense if use them to draw graphs, so export them from EstimateS and load into MS Excel® (or other spreadsheet software).

Click on the 'Export' button at the bottom of the results window, or go to 'Diversity > Export Diversity Stats'. Give the results file a suitable name, such as "LBunut_bats_results.txt".

Find the file "LBunut_bats_results.txt" in My Computer and either right-click and select Open With > MS Excel or drag and drop into an MS Excel window.

The results should now appear in the spreadsheet. Plotting will be simpler if you delete the first few lines, so that the column headings appear in Row 1, then save it in MS Excel's .xls format:

Highlight the first three rows of the spreadsheet and select 'Edit > Delete...' then "Entire row". The column headings should now be in Row 1.

Look at the tab for this spreadsheet near the bottom of the window. It has the same name as the .txt file we imported, "LBunut_bats_results", which is too long for some of the procedures we want to use later. Right-click on the tab name, select "Rename", and change the name to just "LBunut".

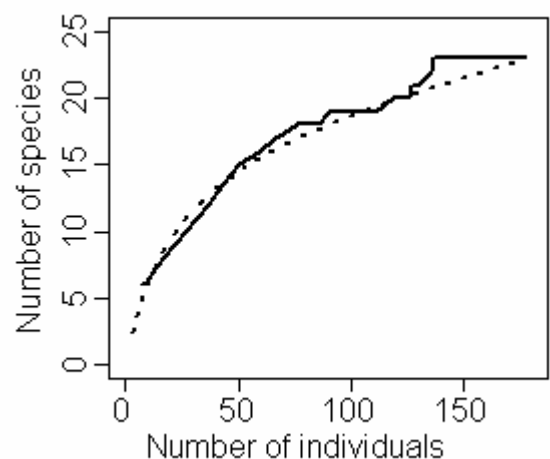
Go to 'File > Save As...', change 'Save as Type:' at the bottom of the dialogue box to "Microsoft MS Excel Workbook (*.xls)".

C. Species accumulation curves

When we looked at the input data, we saw that the number of species observed (*Sobs*) was 22. The first question is, "Have we captured all the species in the population?"

We can get an idea of the answer from the 'collectors curve', a graph of the number of species versus number of individuals as more and more bats are added to the sample. This is the solid line in the figure on the right.

This curve is quite jagged, and the shape depends on the order of the samples. But if the samples are properly independent, the order of collection doesn't matter. We can remove the jaggedness by shuffling the order of the samples many times and averaging the curves obtained, to produce the dotted line in the graph. EstimateS has already done this, shuffling the samples 50 times and taking the mean; the results are in column G [Sobs Mean (runs)] of the spreadsheet. Let's plot it:



In the MS Excel spreadsheet, highlight column B and column G:

Individuals (computed)

Sobs Mean (runs)

(Click on the "B" at the top of column B, then hold down the 'Ctrl' key while clicking on the "G" at the top of column G.)

Start the Chart Wizard by clicking on the toolbar button or using 'Insert > Chart...'

Step 1 : Under 'Chart Type:' select "XY (Scatter)", then in 'Chart Subtype' choose one with "lines without markers". Click 'Next'.

Step 2 : Check that one curve is displayed with the X-axis running from 0 to 200. On the 'Series' tab, you should have one series (Sobs Mean (runs)), with Column B (Individuals) used as the X Values. If all is in order, press 'Next'.

Step 3 : In the 'Titles' tab, put in suitable names, such as:

Chart title: Species accumulation curve

Value (X) axis: No. of Individuals


Value (Y) axis: No. of Species

Press 'Next'.

Step 4: Select "As new sheet:" and name it "Species Accumulation Curve"; click 'Finish'. Save your work (Ctrl-S).

The curve is not exactly jagged, but it isn't really smooth either. More shuffles of the samples would be needed to make it properly smooth. But EstimateS has another way to do this, using an algorithm developed by Mao Chang Xuan to calculate the values we would get for an *infinite* number of randomizations: this is in column C [Sobs (Mao Tau)]. His algorithms also allow us to calculate the confidence interval and those values are in columns D and E. Add these to the graph:

Go to the "Species Accumulation Curve" graph and select 'Chart > Add Data...'

Click on the small  icon at the right-hand end of the 'Range:' box.

Go to the 'LBunut' results spreadsheet and highlight columns B thro E:

- Individuals (computed)

- Sobs (Mao Tau)

- Sobs 95% CI Lower and Upper Bounds

(Click on the "B" at the top of column B, then hold down the Shift key while clicking on the "E" at the top of column E.)

Click on the  icon at the right-hand end of the 'Add data - Range:' dialogue box.

Back in the "Add Data" dialogue box, click 'OK'.

The 'Paste Special' dialogue box opens.

- We want to 'Add cells as' New series, with 'Values (Y) in' columns.

- Make sure that the boxes next to 'Series Names in First Row' and 'Categories (X Values) in First Column' are both checked.

- Click 'OK'.

Three new lines appear in the graph, the middle one being very close to the original *Sobs* curve but much smoother. The outer lines are the confidence limits – let's make them dotted lines, the same color as the Mao Tau curve:

Right-click on the top line and select 'Format Data Series...'. On the 'Patterns' tab, change 'Style:' to a dotted line and 'Color:' to the same colour as the main line (pink in my graph). Click 'OK'

Do the same for the bottom line.

The species accumulation curve we have just plotted is clearly still climbing and showing no signs of leveling out at an asymptote. It looks as though our coverage of species is still rather incomplete.

D. Singletons and uniques

'Singletons' are species which are represented by a single *individual* in the collection. If our sampling has been really thorough, we will have caught all the species there not just once but twice or more. Are there singletons among the bat species trapped at Loagan Bunut?

Go to the LBunut worksheet in MS Excel. Highlight cell D2 (just below the heading Sobs 95% CI Lower Bound), then go to Window > Freeze Panes. Now scroll down and check the last value in the Singletons Mean column (column H).

You'll see there are 8 singletons (out of 22 species in total), so our sampling is not very thorough. Check also the number of 'doubletons': these are species represented by just two individuals in the collection.

If bats of a particular species tend to occur in groups, we'll get a few samples with several bats and lots of samples with none. In this case, the concept of 'uniques' – species which occur in only one *sample* – is more appropriate than singletons. 'Duplicates' are species which occur in just two samples. Check the number of uniques and duplicates in the Loagan Bunut bat data.

Just as for the species accumulation curve, the *trend* in the numbers of uniques and duplicates (or singletons and doubletons) may be more informative than the final numbers. Let's plot the graphs:

In the LBunut spreadsheet in MS Excel, highlight columns B, H, J, L, and N, ie:

- o Individuals (computed)
- o Singletons Mean
- o Doubletons Mean
- o Uniques Mean
- o Duplicates Mean

(Click on the "B" at the top of column B, then hold down the 'Ctrl' key while clicking on "H", "J", "L" and "N".)

Start the Chart Wizard by clicking on the toolbar button or using 'Insert > Chart...'

Step 1 : Under 'Chart Type:' select "XY (Scatter)", then in 'Chart Subtype' choose one with "lines without markers". Click 'Next'.

Step 2 : Check that four curves are displayed with the X-axis running from 0 to 200. On the 'Series' tab, you should have four series, with Column B (Individuals) used as the X Values. If all is in order, press 'Next'.

Step 3 : In the 'Titles' tab, put in suitable names, such as:

- o Chart title: Singletons, doubletons, uniques and duplicates
- o Value (X) axis: No. of Individuals
- o Value (Y) axis: No. of Species

Press 'Next'.

Step 4: Select "As new sheet:" and call it, say, "Singletons, etc"; click 'Finish'. Save your work (Ctrl-S).

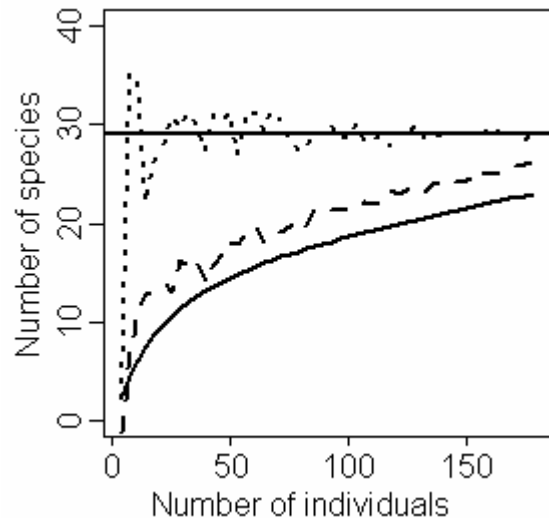
The curves have been smoothed in the same way as the first species accumulation curve we plotted: EstimateS has shuffled the samples 50 times and averaged the results. The shuffling is random, so you will not get exactly the same results each time.

What would you expect to see? When you have only a few samples, you'd expect lots of uniques and singletons; these curves reach a peak at about 30 individuals (8-9 samples) and then start to decline. The duplicates and doubletons reach a peak a bit later, and then also decline. The graph for the Loagan Bunut bats is odd, because the uniques and singletons then start to climb again, and go on climbing. Can you figure out what's happening here?

E. Estimating true species richness

Unless we have collected an enormous amount of data, we will have missed a few rare species. Estimates of species richness based on the number of species observed (*Sobs*) are generally too low. EstimateS provides a number of estimation methods, all based on trying to *extrapolate* from what we know to a situation for which we have no data. You will find a summary of these methods on the web at http://www.wcsmalaysia.org/stats/Biodiversity.htm#rich_estimation.

What are the criteria for a good estimator of species richness? If an estimator gives the correct values for the true richness based on the set of samples we have, that value should not change as we add more samples. Plotted with the species accumulation curve, the ideal estimator would be a horizontal line which meets the species accumulation curve where it levels off – like the solid horizontal line in the graph on the right. In practice, we'd expect the estimate to be very imprecise when we have only observed a small number of individuals, but to settle down to the correct value as we observe more – the dotted line in the graph would be fine. What we often get, however, is an estimate which climbs steadily as we collect more samples, like the dashed line; this leaves us with the same problem – we don't know where it will level off!



EstimateS calculates several estimates of true species richness. Let's plot them and see if any fit our criteria for a good estimator.

In the LBunut spreadsheet in MS Excel, highlight the following columns:

- o B - Individuals (computed)
- o C - Sobs (Mao Tau)
- o P - ACE Mean
- o R - ICE Mean
- o T - Chao 1 Mean
- o X - Chao 2 Mean
- o AB - Jack 1 Mean
- o AD - Jack 2 Mean
- o AF - Bootstrap Mean
- o AI - MMMeans

Start the Chart Wizard and plot a chart as you did for the singletons, uniques, etc. curve. (If you get a message saying "Your formula contains an invalid external reference...", it may be because the name of the results spreadsheet is too long.)

Place the chart in a new spreadsheet and give it a title such as "Species Richness Estimators". Save your work.

The chart is a bit messy with so many lines, and the different colors in the legend are not clear. But if you point to a curve with your mouse, a small box appears with details of the data source.

As you'll see, most of the curves are climbing gradually, running parallel to the *Sobs* curve (the lowest one in the graph). Chao 1 and Chao 2 are climbing more steeply; these are based on the numbers of Singletons and Doubletons (Chao 1) or Uniques and Duplicates (Chao 2) and we saw earlier that these are anomalous. The only one which seems to be reasonably horizontal is "MMMeans" the Michaelis-Menton estimator based on the "Mao Tau" curve for *Sobs*.

The Michaelis-Menton equation describes the progress of an enzyme-catalyzed chemical reaction. There is no theoretical reason to think it might also describe species accumulation, but it does have the right sort of shape, rising quickly at first and then leveling out. The equation is quite simple:

$$S_{obs} = \frac{S_{max} \times n}{B + n}$$

where S_{max} is the maximum number of species – where the curve levels off, n is the number of individuals in the set of samples so far, and B is the number of individuals needed to get half the maximum number of species, ie. when $n = B$, $S_{obs} = S_{max}/2$. The MMMeans column gives us the estimate of S_{max} as we add more samples to the sample set. With all the available samples, the best-fitting curve has $S_{max} = 24.57$ (look at the last value in the MMMeans column) and $B = 35.5$ approximately (look down the Individuals column in the LBunut worksheet to find the value corresponding to $S_{obs} = 24.57/2 = 12.28$ species). Let's plot this curve and see for ourselves if it looks like the species accumulation curve:

Insert a new column next to MMMeans (highlight the column to the right of MMMeans, then use 'Insert > Column') and call it, say, "MM fitted". In the first row of data type

$$=24.57*B2/(35.5+B2)$$

24.57 and 35.5 are the values for S_{max} and B , while the value in cell B2 is the number of individuals, n . Select all the cells in this column from Row 2 to Row 51, then press Ctrl-D to copy the formula to all the cells in the column. Save.

Now plot the 'MM fitted' curve and the Sobs curve – and the MMMeans curve too for good measure: Highlight columns

- o B - Individuals (computed)
- o C - Sobs (Mao Tau)
- o AI - MMMeans
- o AJ - MM fitted

Start the Chart Wizard and plot a chart as before. Place it in a new worksheet with a title such as "MM curve". Save.

The MM fitted and Sobs curves are very close together on the left of the graph, but above about 120 individuals they start to diverge. The "MM fitted" curve will eventually level off at 24.57 species, but the Sobs curve looks as if it is heading higher. The Michaelis-Menton curve does not appear to be a good fit for the part of the curve we are interested in.

None of the estimators implemented in EstimateS seems to perform well for the Loagan Bunut data; we will simply have to collect more data before we can make a sensible estimate.

F. Comparing species richness between sites

When comparing two sites or the same site at two points in time, we can make inferences based on *interpolation*, which is much safer than extrapolation. The process is known as "**rarefaction**".

At Maludam National Park, 81 bats from 11 species were caught. We estimate how many species we would *expect* to find at Loagan Bunut if we only trapped 81 individuals, and this is provided by the smoothed species accumulation curve.

Go to the results spreadsheet for Loagan Bunut, and look down the "Individuals (computed)" column to find the number nearest to 81. Note the number of species – "Sobs (Mao Tau)" – and the upper and lower 95% confidence limits.

With 23 samples we'd catch on average 80.04 individual bats, which is the nearest value to 81. And with 23 samples we'd record on average 16.55 species. This is more than the 11 species we recorded in Maludam NP with a very similar number of bats, so it appears that Loagan Bunut is indeed richer.

Now look at the 95% confidence intervals for *Sobs* for 23 samples: from 10.54 to 22.55. The value of 11 lies within the 95% confidence interval, meaning that finding 11 species in 23 samples from Loagan Bunut is not improbable.

G. Simpson's index: An alternative to species richness

Simpson's Reciprocal index is an alternative to species richness which is less sensitive to the failure to detect rare species and incorporates a correction for small species. The original formula for Simpson's index was:

$$D = \sum \frac{n_i(n_i - 1)}{N(N - 1)}$$

where n_i is the number of individuals of species i in the sample and N is the total number of individuals in the sample. The reciprocal of this (ie. $1/D$) is calculated by EstimateS.

Go back to the LBunut spreadsheet and look for the column headed "Simpson Mean". Note the last value in the column.

Note that EstimateS does not calculate Simpson's index by default, only if you check the box next to "Compute Fisher's alpha, Shannon, & Simpson indexes" on the 'Other Options' tab in 'Diversity > Diversity Settings...' before running the analysis.

The result based on all the samples together is just over 11, which we can think of as indicating that there are 11 common species in Loagan Bunut peat swamps.

Do we have a big enough data set to get a good estimate of Simpson's index, or will it continue to creep up as we collect more bats and discover more species? We'll plot Simpson's index in the same way as the other results:

In the LBunut spreadsheet, highlight the "Individuals (computed)" and "Simpson Mean" columns and plot a graph as we did before for the species accumulation curve and the species richness estimators.

As you will see, the curve rises very quickly to a value above 10 and then soon settles down near the final value of 11. Unlike the species richness estimators, Simpson's index is not much affected by the size of the sample set.

EstimateS calculates two other indices: Fisher's Alpha index and Shannon's index. If you have time, plot these, too, and see how they behave.

H. Finishing off

EstimateS automatically saves the last set of data to be imported and the last set of results generated, which is useful if something goes wrong. But you should carefully keep the input data files and exported results if you wish to preserve them.

Close EstimateS by selecting 'File > Exit' or click on the button in the top right corner of the main EstimateS window.