EXERCISE 2: MULTINOMIAL PROBABILITY AND LIKELIHOOD

Please cite this work as: Donovan, T. M. and J. Hines. 2007. Exercises in occupancy modeling and estimation.
<http://www.uvm.edu/envnr/vtcfwru/spreadsheets/occupancy.htm>

TABLE OF CONTENTS

OBJECTIVES:

- To understand the multinomial distribution and multinomial probability.
- To understand the multinomial maximum likelihood function.
- To determine the maximum likelihood estimators of parameters, given the data.

BACKGROUND INFORMATION

This exercise roughly follows the materials presented in Chapter 3 in "Occupancy Estimation and Modeling." You should work through the Binomial Probability exercise before beginning this one. Click on the sheet labeled "Multinomial" and let's get started.

MULTINOMIAL PROBABILITY

Recall that with the binomial distribution, there are only two possible outcomes (e.g., dead or alive). With a multinomial distribution, there are more than 2 possible outcomes. A common example is the roll of a die - what is the probability that you will get 3, given that the die is fair? In this spreadsheet, we consider only 4 possible outcomes for each trial. For example, in a deck of cards, n = 52 cards, and there are four possible outcomes (hearts, clubs, diamonds, and spades), and we know that p = 0.25 for each kind of card if the deck is complete. As another example, perhaps you are studying a population of plants in which there are four different kinds of phenotypes (y1, y2, y3, and y4), and each phenotype has a certain probability of being included in your study population.

The spreadsheet set up is similar to the binomial sheet.

	F	G	H	I
3	**INPUTS**			
4	**n =**	**50**		
5	**y1 =**	**30**	**p1 =**	**0.1**
6	**y2 =**	**5**	**p2 =**	**0.1**
7	**y3 =**	**10**	**p3 =**	**0.1**
8	**y4 =**	**5**	**p4 =**	**0.7**
9	**sum =**	**50**	**sum =**	**1**
10				
11	$f(y_i \mid n, pi) = \binom{n}{y_i} p_1^{y1} p_2^{y2} p_3^{y3} p_4^{y4}$			
12				
13	**Multinomial Probability =**			0.0000

Enter the total number in the population (trials) in cell G4. Enter the number of times out of n that you observed each outcome in cells G4:G8. For example, assume you sample 50 plants (n = 50), of which 30 are y1, 5 are y2, 10 are y3, and 5 are y4's. Enter 50 in cell G4, 30 in cell G5, 5 in cell G6, 10 in cell G7, and 5 in cell G8. These cells are named _y1, _y2, _y3, and _y4 respectively. Note, the sum of the outcomes computed in cell G9 must equal the value you entered in cell G4.

Let's start by assuming that you DO know the probabilities for obtaining each outcome independently. That is, you KNOW that true proportions of each phenotype in the population. These go in the light blue cells I5:I8, and they must sum to 1. Note that these cells are also named; _p1, _p2, _p3, and _p4. Suppose p1 = p2 = p3 = 0.1, and p4 = 0.7, such that phenotype 4 is much more common than the other phenotypes.

	F	G	H	I
3	**INPUTS**			
4	**n =**	**50**		
5	**y1 =**	**30**	**p1 =**	**0.1**
6	**y2 =**	**5**	**p2 =**	**0.1**
7	**y3 =**	**10**	**p3 =**	**0.1**
8	**y4 =**	**5**	**p4 =**	**0.7**
9	**sum =**	**50**	**sum =**	**1**
10	$f(y_i \mid n, pi) = \binom{n}{y_i} p_1^{y1} p_2^{y2} p_3^{y3} p_4^{y4}$			
11				
12				
13	**Multinomial Probability =**			0.0000

Given these probabilities, the probability of obtaining the field results in cells G5:G8 (the number of plants of each phenotype) can be computed with the multinomial probability function, shown in the purple box.

$$f(y_i \mid n, pi) = \binom{n}{y_i} p_1^{y1} p_2^{y2} p_3^{y3} p_4^{y4}$$

For example, if p1 = p2 = p3 = 0.1, and p4 = .7 (cells I5:I7 = 0.1 and cell I8 = 0.7), the probability of getting the 30 y1, 5 y2, 10 y3, and 5 y4's is given in cell I13. It has a very low probability (because you'd expect more y4 outcomes than you actually observed).

Enter different observations for y1 to y4 in cells G5:G8 - make sure they sum to whatever is entered in cell G4 - and examine the multinomial probability in cell I13.

MULTINOMIAL MAXIMUM LIKELIHOOD

But we often don't know what p1, p2, p3, and p4 are. Instead we know the total number of trials (in this case, plants sampled) and the total number of each outcome. In other words, we don't know what values go in cells I5:I8, we only know

what's entered in yellow cells G5:G8 and what's computed in cell G9. Our goal for this particular spreadsheet, therefore, is to find maximum likelihood estimates of p1, p2, p3, and p4, given the data.

So we use likelihood procedures again. The formula is given in the yellow box. How does this formula differ from the formula in the pink box?

$$L(p_i \mid n_i, y_i) = \binom{n}{y_i} p_1^{y1} p_2^{y2} p_3^{y3} p_4^{y4}$$

Given the data (y's and n), the spreadsheet computes the likelihood of observing the data under different scenario's of p1, p2, p3, and p4 in cell H18. The log likelihood is computed in cell H19. Play around with cells I5-I8 and note how the likelihood changes. Make sure that your probabilities sum to 1.

Since we don't know what p1, p2, p3, or p4 are, the goal is to compute the likelihood (cell H18) or log likelihood (cell H19), given the data, and then to find those values of p1, p2, p3, p4 that maximize the likelihood. There are several ways to do this. We could enter various combinations of p1, p2, p3, and p4 in cells I5:I9, and keep track of the likelihood values and which combinations give the highest likelihood value. This is the hunt and peck method. You can also systematically look at all combinations of p1, p2, p3, and p4 - as shown in a grid of cells in columns M through P (cells M4:P8439). (I generated these combinations in SAS - there are ~900 different combinations here. You can examine a lot more combinations if you'd like but it takes the spreadsheet longer to calculate the cells). Note that column P (p4) could be computed because the sum of the four probabilities must equal 1 (p1+p2+p3+p4 = 1).

Then, for each combination, we compute the multinomial likelihood, given in column K (cells K4:K921). The log likelihood is computed in column L (cells L4:L921). For example, cell L4 has the equation =y_1*LN(M4)+y_2*LN(N4)+y_3*LN(O4)+y_4*LN(P4), and generates the log likelihood estimate when p1 = 0.05, p2 = 0.05, p3 = 0.05, and p4 = 0.85. Now find the maximum value in either column. The log-likelihood value is computed in cell H20 with a MAX function. The estimates of p1, p2, p3, and p4 at this maximum are the maximum likelihood estimates. These are computed in cells J5:J8 with a VLOOKUP function. Note that it might take EXCEL a bit of time to recalculate your cells, depending on the speed of your computer. Also note that in some cases there may be more than one combination of p's that are maximized. If more than 1 combination of p's in this spreadsheet yield the same maximum likelihood, the count in cell H21 will be greater than 1...try some other values for n or the y's.

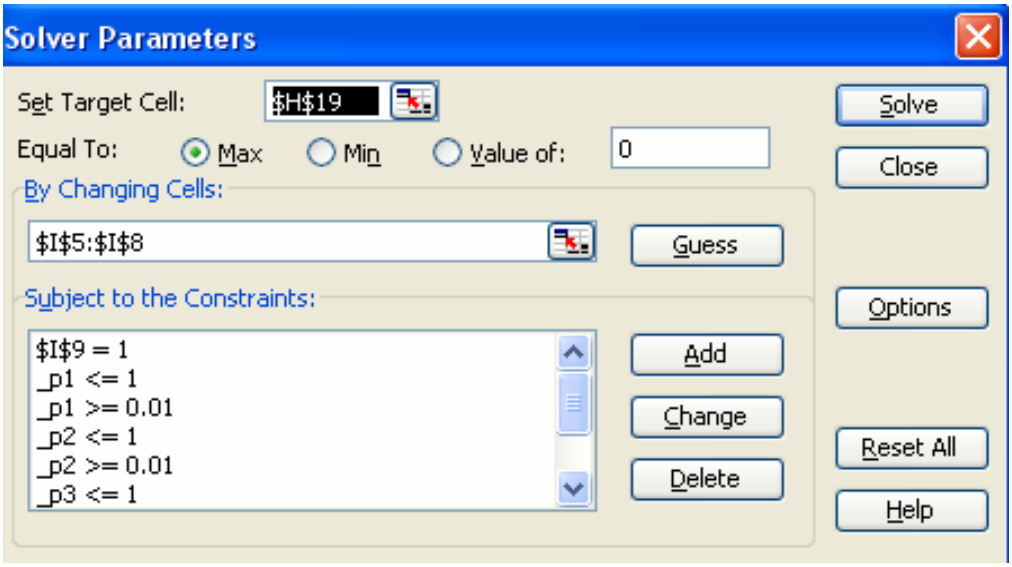
GRAPHING THE MLE'S

Imaginary step! Graph the likelihoods as a function of various p's. If you could graph 3 axes (p1, p2, and p3) and plot the log likelihood as a function of the p's, you should see the same type of hump we saw with the binomial likelihood function. You don't need to graph p4 because p4, by definition, is 1 - p1 - p2 - p3. The graph gives an indication of which combinations of p1, p2, p3, and p4 yield the highest likelihood values.

USING SOLVER TO FIND THE MLE's

Rather than going through all combinations of p1, p2, p3, and p4 by brute force, we can also use the Solver to find the MLE across the 4 p parameters for us. The Solver is a numerical optimization program within Excel. We'll use the Solver in

future spreadsheets too, so make sure Solver is installed (if not, go to Tools | Add-ins, and select the Solver box). To access Solver, go to Tools | Solver. A dialogue box should open. You want to set the target cell (the log likelihood given in cell H19) to a maximum, so enter H19 and select the "maximize" radio button. You'll find this maximum by changing cells I5 to I8 (p1, p2, p3, p4). You'll also need to add some constraints: each of the the p's can take on values between 0.01 and 1, and the sum of the p's must be 1 (cell I9 must equal 1). Note that the image below does not show all of the constraints.



Once the constraints have been entered, press Solve and the Solver will try various combinations of p's in cells I5:I8 to find the maximum log likelihood. The Solver will go through an iterative process and pause on results that it obtains for each iteration - press Continue until a final solution has been reached. Compare the Solver estimates with the "grid search approach" in cells J5:J8. If Solver computes negative estimates for some reason, run it again. NOTE: The Solver sometimes crashes, especially when there are few observations of a particular event and the p estimates for that event are very low. Later on we'll add something to keep it from crashing.

SUMMARY

That wraps thing up for this exercise. Make sure you have a good understanding of what multinomial maximum likelihood is before proceeding to any of the other spreadsheet exercises. The multinomial maximum likelihood function is the workhorse for ALL of the occupancy modeling exercises presented in the book, "Occupancy Estimation and Modeling." If you don't truly understand the multinomial maximum likelihood function, you won't truly grasp what your results indicate or how your parameters were estimated. As we go through occupancy models in more detail, hopefully the application of multinomial maximum likelihood will become clear.