EXERCISE 1: BINOMIAL 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>

OBJECTIVES

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

BINOMIAL DISTRIBUTION

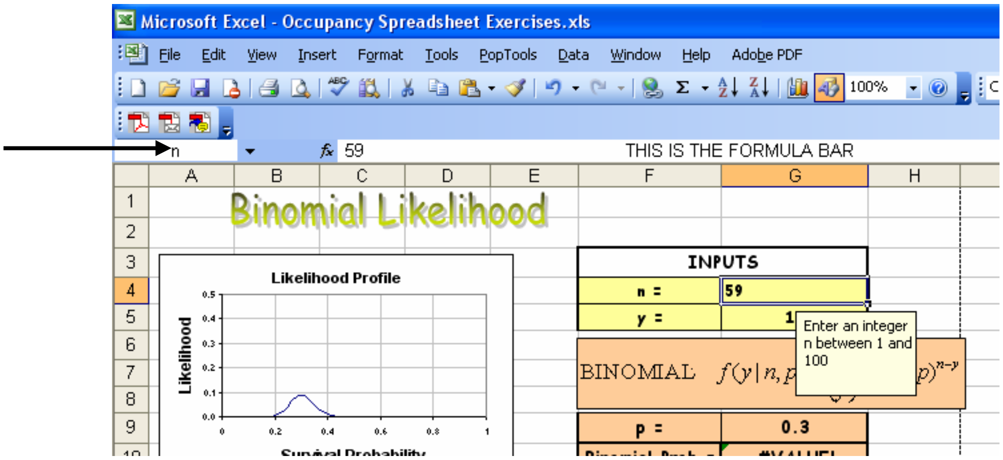
This exercise roughly follows the materials presented in Chapter 3 in “Occupancy Estimation and Modeling.” Click on the sheet labeled “Binomial” and let’s get started. The binomial distribution is widely used for problems where there are a fixed number of tests or trials (n) and when each trial can have only one of two outcomes (e.g., success or failure, live or die, heads or tails). The formula is written in the orange box below and on the spreadsheet:

$$\boxed{\text{BINOMIAL:} \quad f(y \mid n, p) = \binom{n}{y} p^{y} (1-p)^{n-y}}$$

The number of successes is usually denoted as y, and the probability of success is usually denoted as p. A typical example considers the probability of getting 3 heads, given 10 coin flips and given that the coin is fair (p = 0.5). The left side of the binomial probability function is written *f*(3|10,0.5), where the vertical bar | means “given” and is read, “the probability of getting 3 heads, *given* 10 coin flips and the probability of a head (success) is 0.5.” Let’s break the right hand side of the binomial probability function into pieces. The portion $\mathbf{p^{y}}$ and $\mathbf{(1-p)^{n-y}}$ gives p (the probability of success, or heads) raised to the number of times the success occurred (y) and 1-p (the probability of a failure, or tails) raised to the number of times the failures occurred. But if you flip a fair coin 10 times, there are many ways you could end up with three heads. For instance, the first three tosses could be heads and the rest could be tails (HHHTTTTTTT). Or the first seven could be tails and the last three could be heads (TTTTTTTHHH). Or you could alternate getting heads and tails (e.g., THTHTHTTTT). The portion of the binomial probability function in brackets is called the binomial coefficient, and accounts for ALL the possible ways in which three heads and seven tails could be obtained.

Another example considers the probability of 15 sites are occupied by a species of interest out of a 59 total sites, given p = 0.3 (the probability of survival). This example assumes that detection probability (the probability of detecting a species that occurs on site, given the site is occupied) is perfect and can be ignored. The left side of the binomial probability function would be written: $f(15|59,0.3)$.

The binomial probability can easily be computed in a spreadsheet environment. In the spreadsheet, you'd enter the following data: **n** = 59 (cell G4), **y** = 15 (cell G5), and **p** = .3 (cell G9). Click on cell G4, and then look to the left of the formula bar…you'll see that this cell has been named **n**. Naming cells is a spreadsheet option that we'll sometimes use in this book to help clarify formulas. (To name a cell, just click on the cell and go to Insert | Name | Define. Or just click on cell and start typing the name where the cell address is listed to the left of the formula bar.) Similarly, cell G5 is named **y**, and cell G9 is named **p**.



Microsoft Excel - Occupancy Spreadsheet Exercises.xls

n | 59 | THIS IS THE FORMULA BAR

Binomial Likelihood

INPUTS

n = 59

y = 1

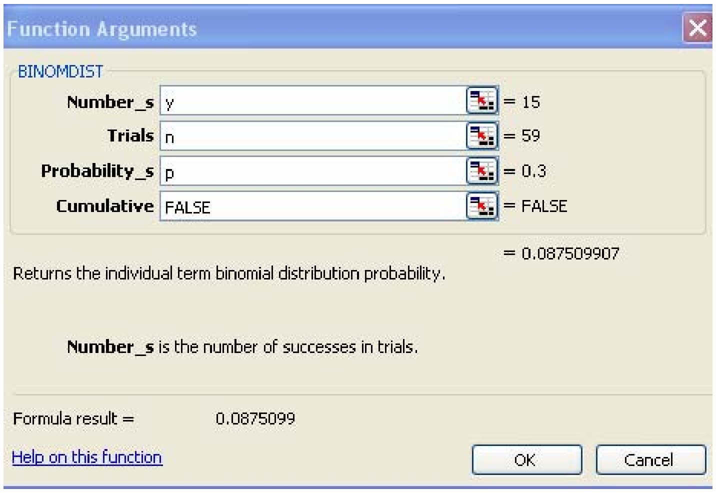
Enter an integer n between 1 and 100

BINOMIAL $f(y|n,p$ … $p)^{n-y}$

p = 0.3

Binomial Prob = #VALUE!

Given the entries for n, y, and p, you can use the canned Excel function called **BINOMDIST** to compute the probability of getting 15 occupied sites out of a population of 59 (cell G10), which is 0.0875099. Click on cell G10 and you'll see the equation: **=BINOMDIST(y,n,p,FALSE)**, where BINOMDIST is the name of the Excel function, and the information in the parentheses are function's arguments. This particular function has four arguments:



BINOMDIST

Number_s	y	= 15
Trials	n	= 59
Probability_s	p	= 0.3
Cumulative	FALSE	= FALSE

= 0.087509907

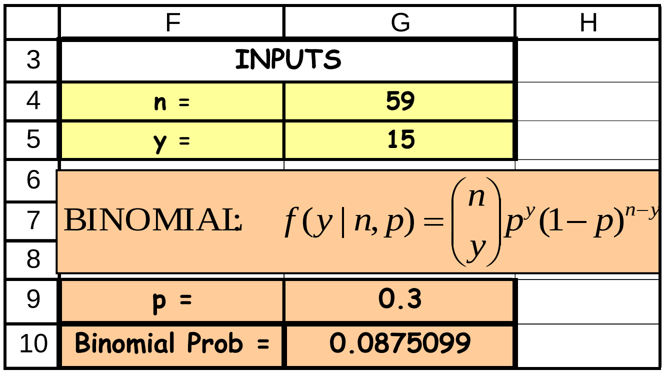
Returns the individual term binomial distribution probability.

Number_s is the number of successes in trials.

Formula result = 0.0875099

Help on this function

In this example, the *names* of the cells used as arguments appear instead of the cell addresses (E.g., Number_s is the number of successes, and is entered in cell G5, which is named y, so y appears instead of G5). Given the arguments of the function, Excel returns the binomial probability in cell G10:

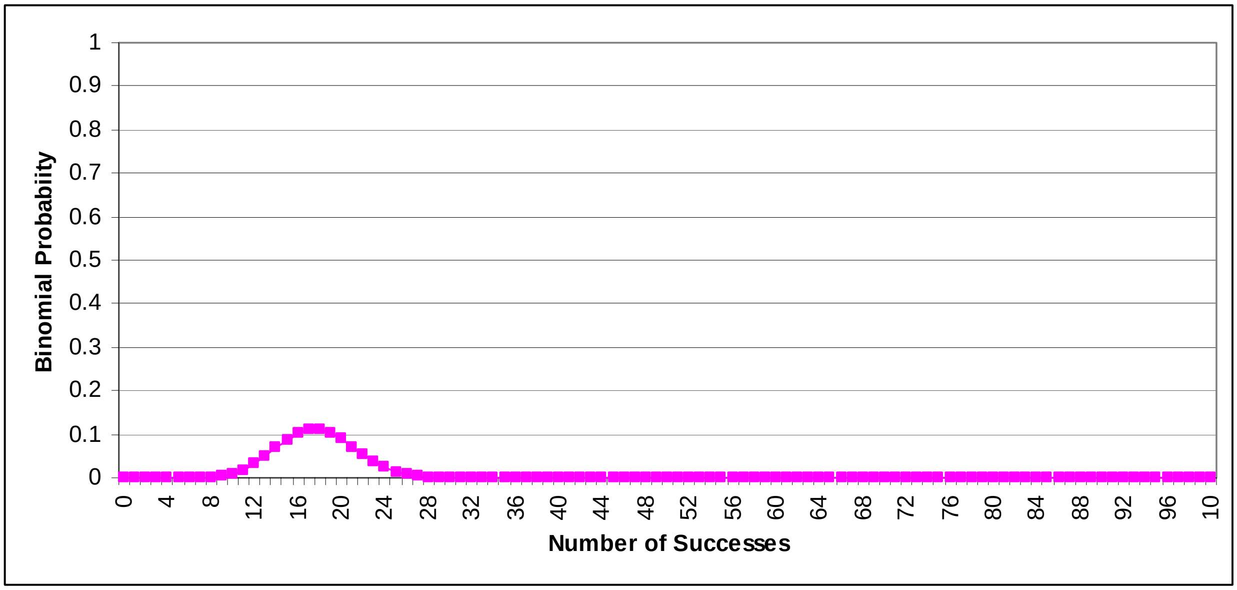


	F	G	H
3	**INPUTS**		
4	**n =**	**59**	
5	**y =**	**15**	
6			
7	BINOMIAL:	$f(y \mid n,p)=\binom{n}{y}p^y(1-p)^{n-y}$	
8			
9	**p =**	**0.3**	
10	**Binomial Prob =**	**0.0875099**	

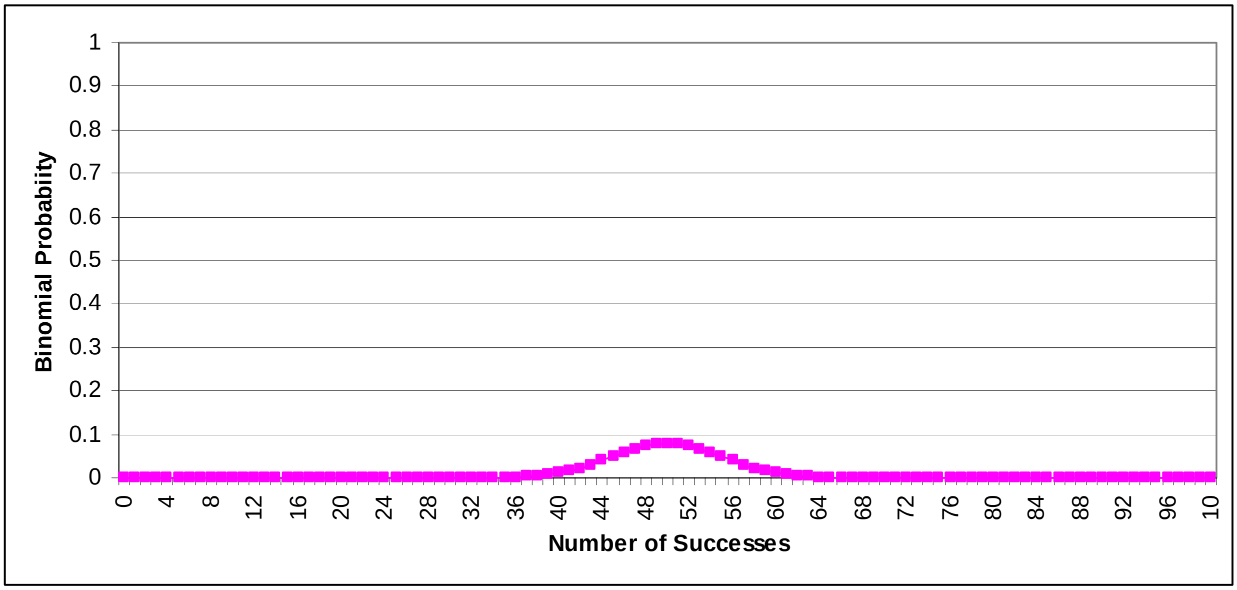
That is, the probability that 15 sites are occupied by the species of interest, given that 59 sites are surveyed and the probability of occupancy is 0.3, is 0.0875. Try different entries in cells G4:G5,G9 to get a handle on binomial probability. But keep n under 100 or your spreadsheet will have computational troubles (unless you use scientific notation)!

GRAPHING THE BINOMIAL DISTRIBUTION

It's useful to graph the binomial probability to see the entire probability density function. In cell F24:F124, we entered the numbers 0 through 100 to represent the total possible number of successes (assuming a maximum of 100 trials). In cell G24, we entered the equation =BINOMDIST(F24,n,p,FALSE), which is the same as before except that we reference cell F24 for the number of successes instead of cell G5. This formula is then copied down to cell G124, and the results are graphed.



This graph is the binomial probability function for n = 59, p = 0.3. You can see that, given the data, it would be very unlikely to have few sites or more than 28 sites occupied by the species of interest, given the data. The binomial probability function changes as n and p change. For example, below is the function for n = 100 and p = 0.5:



ASSUMPTIONS OF THE BINOMIAL DISTRIBUTION

Two major assumptions of the binomial distribution are that the trials are independent, and probability of success is constant throughout the experiment. (Note: sometimes these assumptions are violated. If you flip a penny, the outcome of the next flip will be completely independent of the outcome of the first flip. But animals are not pennies. Pair bonds, home range size, dispersal, and family associations are examples of how the occupancy of one site can be linked to the occupancy state of another site, resulting in extra binomial variation. How to deal with this problem is covered later.)

BINOMIAL LIKELIHOOD

But we often don't know p in field biology. What we DO know is the number of successes or sites that are occupied (y) and the total number of trials or sites (n). The goal of this spreadsheet is to show how likelihood procedures can be used to estimate p, *given* n and y, with maximum likelihood procedures. Just to keep things clear, from this point on let's assume that p is the probability of occupancy, n is the number of sites in a population, and y is the actual number of sites that are occupied by the species of interest.

In cell G4 enter n = 100, or the number of sites in the population. In cell G5 enter y = 50, or the number of sites where the species of interest was found. Note that y must be less than or equal to n. We don't know p (cell G9), so you can either ignore this cell or delete the values in it. Given those data, the goal is to find the maximum likelihood estimate (MLE) of occupancy, or p. This equation is shown in the green box.

$$\text{LIKELIHOOD:} \quad L(p \mid n, y) = \binom{n}{y} p^{y}(1-p)^{n-y} = \frac{n!}{y!(n-y)!} p^{y}(1-p)^{n-y}$$

Note that the binomial coefficient can be written in two ways:

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

OK, how does the binomial likelihood function differ from the binomial probability function in the orange box?

$$\text{BINOMIAL:} \quad f(y \mid n, p) = \binom{n}{y} p^{y}(1-p)^{n-y}$$

You should see that the information to the left of the equal sign differs between the two equations, but the information to the right of equal sign is identical. The binomial probability function estimates the probability of getting y successes, given n and p (where again the vertical line | means "given"), while the binomial likelihood function estimates the probability of p, given n and y.

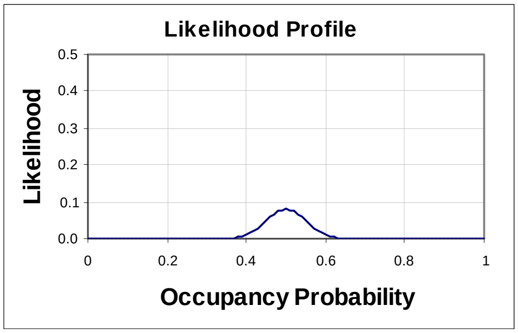
The spreadsheet is set up to compute the likelihood estimate for a variety of p estimates. In column K, cells K4:K104, we let p vary from 0 to 1 in increments of 0.01. For each p, the likelihood is computed in column L (cells L4:L104) - the formulae in these cells follow the formula outlined in the green box.

$$\boxed{\text{LIKELIHOOD:} \quad L(p \mid n, y) = \frac{n!}{y!(n-y)!} p^{y}(1-p)^{n-y}}$$

Prove this to yourself by clicking on cell L7 (for example) and examining the formula in the formula bar. The formula in cell L7 is =FACT(n)/(FACT(y)*FACT(n-y))*K7^y*(1-K7)^(n-y) . (Note: You could also have used the COMBIN function instead of the FACT function. FACT(n)/(FACT(y)*FACT(n-y)) can be written as COMBIN(y,n). Also, if a cell is set to the scientific format, Excel can handle larger numbers).

GRAPHING THE MLE

For any given n and y, we can graph the likelihood profile, which tells us the likelihood value for each and every p estimate. Examine the shape of the likelihood profile below.



On the x axis is p, which ranges from 0 to 1. On the y axis is the likelihood value. The graph shows the range of likelihood values possible, given the data. The probability value where this graph peaks is the maximum likelihood estimate, or MLE; it shows where the likelihood is greatest.

Graphing is one way to find the MLE. There are other ways too. For example, Excel finds the maximum likelihood value with a MAX function and displays it in cell H14. The corresponding probability associated with this maximum likelihood is computed in cell G15 with a VLOOKUP function. Another way to find the MLE is to solve for it analytically. The analytic solution is found by setting the first derivative of the likelihood equation to 0, and solving for p. This means (conceptually) that for every point on the graph, you estimate its tangent line and find the tangent line whose slope is 0 – that's the top of the curve. The analytic solution to the maximum likelihood estimate is computed in cell H15. Cells G15 and H15 should match each other (or be very close).

THE LOG LIKELIHOOD FUNCTION

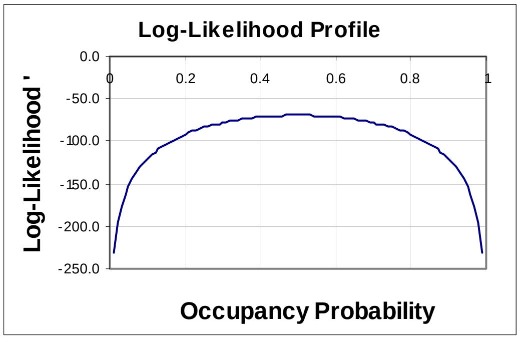
The log likelihood function is analytically easier to work with than the likelihood function because it drops the binomial coefficient. The log-likelihood is computed in column M (cells M4:M104), and the formulae in these cells follow the formula outlined in the blue box. Note that the equal sign has been replaced by a symbol which indicates that the log likelihood of p given the data is proportional to yln(p)+(n-y)ln(1-p).

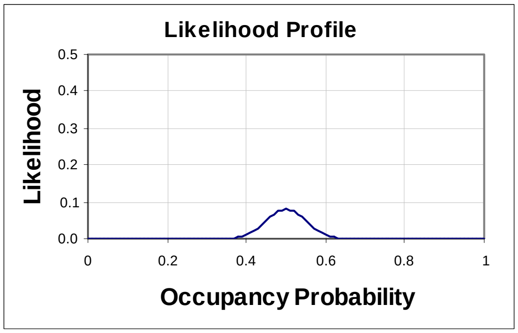
$$\boxed{\text{LOGLIKELIHOOD:}\quad \ln(L(p\,|\,n,y)) \propto y\ln(p)+(n-y)\ln(1-p)}$$

Note also that the log-likelihoods are negative, and that Excel returns a #NUM! error message for p = 0 and p = 1.

The maximum log-likelihood is computed in cell H18 with a MAX function. The corresponding p value for this maximum is computed in cell G19 with a VLOOKUP function, and the analytic solution is computed in cell H19. These two values should be the same.

A graph of the log-likelihoods against p is also shown (labeled the log-likelihood profile). Notice that the shape of the log-likelihood function is an upside-down U, which looks comparable to the probability density function.

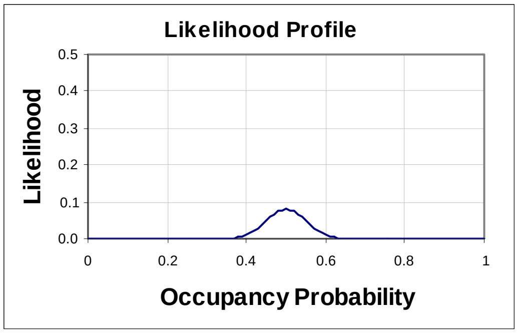


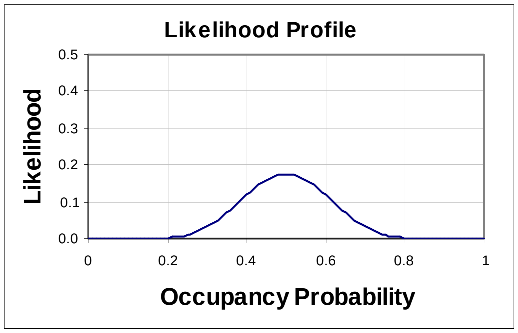


The peak of this graph is the MLE. Compare the MLE for the likelihood profile and the log-likelihood profile - they should be the same.

PRECISION OF THE MLE ESTIMATOR

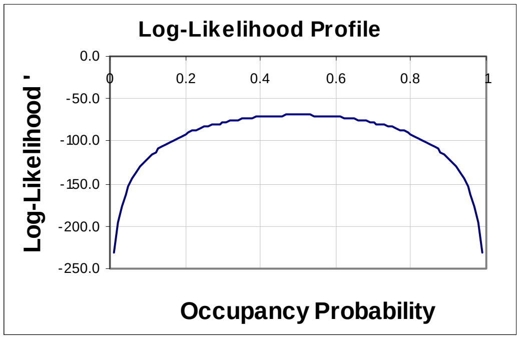
Now that we have found the MLE, given the data, we need to know something about the precision of this estimate. The sampling variance of the MLE is directly related to the curvature of the likelihood function or the log likelihood function at its maximum. Compare the MLE and shape when n = 100, y = 50 (left) with n= 20, y = 10 (right):

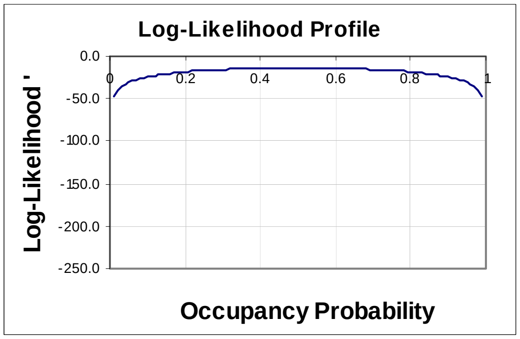




Notice that both estimate the MLE as p(hat) = 0.5. But notice also that when the sample size is greater, the variance around the MLE is tighter (left diagram) than when the sample size is smaller. Take home message: bigger sample sizes produce more precise MLE's.

Here are the log likelihood profiles for n = 100, y = 50 (left) and n = 20, y = 10 (right).





Obviously the graph on the left has more curvature than the graph on the right (where the right graph is the lower sample size). When the log likelihood profile has a "good" curvature, it suggests a lower variance in the MLE estimate compared to a profile that has "poor" curvature. In the graph on the right, there are many values of survival probability that generate almost the same log-likelihood value, suggesting that you should not be overly confident that the MLE is actually 0.5. Although both graphs have an MLE of 0.5, our "confidence" in this estimate is much lower when the sample size is smaller. In a nutshell, getting the MLE is only the first step of the analysis – you also need to estimate the variance of that estimate before you can make inferences.

Here's a clip from FW663 lecture notes (http://www.warnercnr.colostate.edu/class_info/fw663/):

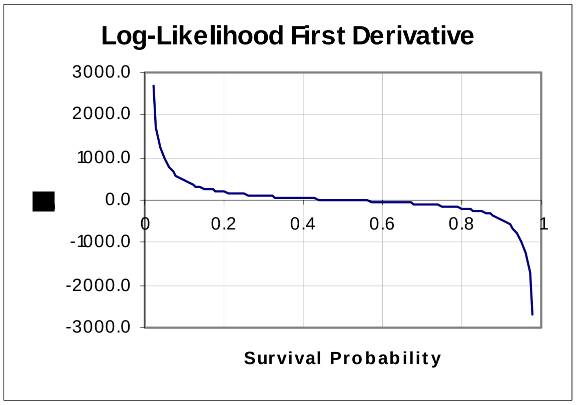
"If the log-likelihood function is fairly flat, this implies considerable uncertainty and this is reflected in large sampling variances and standard errors, and wide confidence intervals. On the other hand, if the log-likelihood function is fairly peaked near its maximum point, this indicates some values of *p* are relatively very likely compared to others. There is some considerable degree of certainty implied and this is reflected in small sampling variances and standard errors, and narrow confidence intervals. So, the log-likelihood function at its maximum point is important as well as the shape of the function near this maximum point."

How exactly do we quantify the degree of curvature of a function? We can determine "curvature" at the MLE by examining the second derivative of the log-likelihood function - this tells us how rapidly the likelihood function accelerates - the more rapidly it accelerates, the lower the variance (the sharper the log-likelihood function). Fisher showed that the sampling variance of the MLE is the negative inverse of the second derivative, evaluated at the MLE.

Let's walk through the calculations, starting at the beginning. For any given point on the curve, the derivative is the slope of the line tangent to that point on the curve. This is calculated in column O (cells O4:O104) with a 3 point estimate. Be sure to look at the formula - it's a slope equation. In this case, we estimate the slope at a particular point on the likelihood surface by examining the values of the two points surrounding it. If you're drawing a blank right now, it might be useful to recall that for two points, $(y_2-y_1)/(x_2-x_1)$ provides a slope estimate. For example, click on cell 06 and you'll see a formula =(M7-M5)/(K7-K5). This gives us the slope of the log likelihood function when p = 0.02. The answer is 2695.1, which is rise over run. This means the function is almost vertical at this point. Let's try another point. What is the derivative of the log likelihood function when p = 0.5? Click on cell 054 and you should see the formula =(M55-M53)/(K55-K53). The answer is 0, indicating that the slope of a line tangent to this point on the function = 0, in turn indicating that it is the top of the function.

A graph of these slopes across all points is a function itself (scroll down to a graph labeled Log-Likelihood First Derivative), and is technically called Fisher's score function. If you have studied calculus, you wouldn't use the spreadsheet approach to find the slope of a tangent line at individual points; instead you'd find the derivative of the function analytically, which is an equation that describes the change in the likelihood across all points taken together. This equation is given in Column R. The score is a vector of first partial derivatives solved analytically. A plot of the first derivative over different values of p is shown below, and is technically written as:

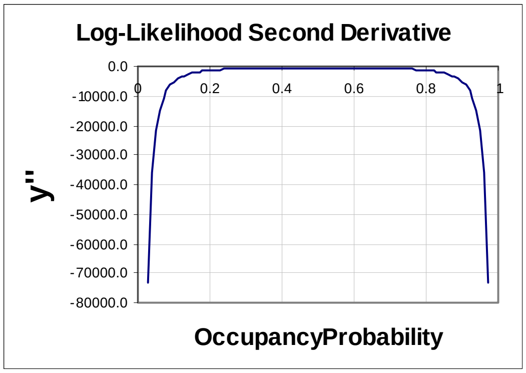
$$\text{Score Function} = y' = \frac{\partial \text{LogL}(\text{p} \mid \text{y})}{\partial \text{p}}$$



When the score function is positive (y' > 0), it means the log-likelihood function is going up, and when it's negative (y' < 0), it means the log-likelihood function is going down. When this function is 0 (y'= 0), it means the slope at a given point on the original log-likelihood graph is 0, which happens at the peak of the curve, or the MLE. Look at the graph labeled First Derivative. You should see that where y' = 0, x is the MLE.

Play around with different values in cells G4 and G5, and examine the likelihood function, the log-likelihood function, and the first derivative of the log-likelihood function. Look at the graphs and determine the MLE.

Now we're ready for the second derivative of the log-likelihood function (which according to Cooch and White is called the Hessian). It is just the slope of the plot of the first derivative values (or the slope of the slope of the original function), and is computed in cells P4:P104 with another 3 point estimate. A graph labeled Log-Likelihood Second Derivative is shown.



So now that we have the second derivative estimates, we can compute the sampling variance as -1 * the inverse of the second derivative (evaluated at the MLE). This formula is entered in cells Q4:Q104. The variance of the log-likelihood function is computed in the spreadsheet with a VLOOKUP function in cell G20. The spreadsheet looks up the MLE, and returns its associated variance. The analytic formula is computed in cell H20. Cell G20 should match cell H20 – but they may not match exactly because the 3-point approach is less accurate than the analytic approach. Columns R and S are the analytic solutions of the first and second derivatives:

$$Likelihood' = \text{score function} = \frac{y}{p} - \frac{n-y}{1-p}$$

$$Likelihood'' = \text{Hession} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2}$$

The purpose of all of this is to simply provide an intuitive look behind how variance is estimated for a parameter. We'll return to this topic, as well as profile likelihood confidence intervals, in the single-species, single-season occupancy model.