Ch. 3

Maximum likelihood estimation[1]

"Every side of a coin has another side."
-- Myron Scholes

Questions to ponder:

- *What is a maximum likelihood estimator?*
- *No, seriously, what is a maximum likelihood estimator?!*

An introduction

Maximum likelihood estimation is a concept that is central to everything that we will discuss in this primer. So, there are some very good reasons for a student to learn the basics of maximum likelihood estimation. First and foremost, it is a phrase that will impress almost all of your colleagues, if you can use it in an appropriate sentence at the appropriate moment!

Unfortunately, maximum likelihood estimation is not well-understood by many who fear the quantitative explanations given by many textbooks. In fact, one of us (LP) was 90% of the way through his first parameter estimation course as a Master's student before he asked his officemate (a PhD student) if he could explain the phrase to him—they were using it in class all of the time, and he didn't know what it meant. Surprisingly, the PhD student couldn't explain it either. So, LP and his colleague sat down together, and looked for easy-to-understand explanations. Here, we'd like to take you through the same learning process that LP went through as a student. Maximum likelihood estimation is the most widely used method of parameter estimation—you need to learn the basics to appreciate the estimates you will eventually obtain.

We promise to take it slow.

If we break the phrase down into individual words, it may help with our understanding. Imagine that you have a sample of mark-recapture or survey data. And, imagine that you are trying to estimate the value of a parameter, such as density or the probability of survival. So, you could ask yourself, *"Given my data, what is the most likely estimate for the parameter*?"

---

[1] *With thanks for content to Therese Donovan.*

If you asked that question, you have just framed the concept of a maximum likelihood estimator. Using mathematics, **maximum likelihood estimation** is a method to **estimate the most likely value of a parameter, given a sample of data.**

Remember that we are operating "in the dark", so to speak. We do not know the true value of a parameter, but we do have our field observation data (capture records of marked animals, distance estimations of detected animals, number of sites occupied etc.). And, we can construct simple models of probabilities that we postulate to be true (these are our assumptions). So, the use of a maximum likelihood estimation method falls under the category of inductive reasoning: *if A and B are true and our sample is representative of our population, then it appears that the value of the parameter for this population is most likely X.*

Statisticians are good at helping us operate in the 'dark', without full knowledge of a population. If you value the input from smart statisticians, you'll be happy to know that maximum likelihood estimation methods are well-accepted, and can be traced back to Sir Ronald Fisher (*Fisher has been credited with providing the foundations for modern statistical science, as well as providing ground-breaking ideas for evolutionary biology*). Statistically, maximum likelihood estimators are unbiased, especially for large samples, and the variance provided by the estimator is minimized, especially for large sample sizes. Furthermore, maximum likelihood estimators are approximately normally distributed, and even non-statisticians are aware that we like to work with normal distributions when we conduct statistical tests.

Statisticians tell us that the MLE method is easy to apply—we will let you decide for yourself after you finish this chapter. *If you find it easy, you may have found your calling as a quantitative ecologist!*

We can also make a list of reasons that are important to us, as biologists who work with quantitative methods, to become familiar with maximum likelihood estimation:

- Because maximum likelihood estimation is based on **likelihoods**, or **probabilities**, *the method is generally intuitive to us*—that is, we can make sense of it. In complex situations, we may agonize over the equations for probability statements (likelihoods), but the method allows us to 'work it out' in our brains.
- Likelihood methods are useful for parameter estimation, and we will also see (in the next chapter) that *likelihood theory is useful when we are conducting model comparison* with methods such as Akaike's Information Criterion.
- The use of maximum likelihood estimation *allows us to easily obtain estimates of variance* for our parameters.

Binomial coefficients

Most of the estimators that we will employ in this primer are based on a basic concept—we can establish probabilities associated with samples in a **binomial trial**. That is, in a trial for which there are only two possible outcomes (success/failure, heads/tails, alive/dead, emigrated/not emigrated, etc.), we can use basic mathematical facts to establish probabilities for the occurrence of an event (e.g., an animal living, or a head of a coin during a toss). So, we will begin our study of maximum likelihood estimation with the concept of the binomial coefficient to incorporate

probabilities, or likelihoods. *Note: in latter chapters we will use* ***multinomial*** *coefficients when we have more than two possible outcomes such as when animals may be located in >2 areas during a survey or capture event.*

We use the expression of the binomial coefficient to calculate the number of ways, or the number of different combinations, a sample size of *n* can be taken from a population of size *N*. We express the binomial coefficient as:

$$\binom{N}{n}$$

And, we read it *"Big-N choose little-n"*.

For example, given a group of 3 individuals (N =3), how many combinations of two individuals (n=2) can be found? Generally, the formula for a binomial coefficient tells us that that we can calculate our answer as:

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

As a review, N! is read "N factorial". N! is the product of N*(N-1)*(N-2)*(N-3)*… until the last difference is 1.

Thus: 5! = 5*4*3*2*1=120.
And: 10! = 10*9*8*7*6*5*4*3*2*1 = 3,628,800.

So, to utilize the factorial statements in our example,

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1(1)} = 3$$

Therefore, there are 3 combinations of two individuals that can be drawn from a group of 3 individuals. We can check the math with some visual logic, as our population is small and manageable. Let us give identifying letters to our 3 individuals: A, B, and C. We can see that it is possible to draw the following combinations of letters:

AB
AC
BC

So, there are three pairs or combinations, just as predicted. *Note that in this situation, the order of the letters does not matter. "AB" is the same combination as "BA".*

Let's do another example that may be a little more representative of something that we'll want to do for parameter estimation. Let's say that we are going to have 20 tosses of a coin—20 trials of

a coin flip. How many combinations of coin tosses would result in exactly 5 heads (and thus, 15 tails)? We know we could start to write these out—the first three might be:

HHHHHTTTTTTTTTTTTTTT
THHHHHTTTTTTTTTTTTTT
TTHHHHHTTTTTTTTTTTTT
TTTHHHHHTTTTTTTTTTTT

You can see that this is going to take a while! And, the heads don't have to be all in a row, either…so, we better use math rather than just writing it out. Math can save us a lot of time!

We take our general formula and put in our values for N=20 and n=5. Give it a try. You should find that there are 15,504 combinations of heads and tails, from first toss to the 20th toss, that result in 5 heads. Aren't you glad we used math instead of writing them out?!

To make this 'biological', we could think of a radio-telemetry project with 20 bighorn sheep (*Ovis canadensis*). If we can imagine an experiment ending with 5 of the sheep alive after a year, how many combinations of sheep (sheep #1, #6, #7, #10, and #18, for example) could we have that allow 5 sheep alive?

Figure 3.1: *Radio- and ear-tagged male bighorn sheep in Nebraska (photo by Kelly Powell, used with permission).*

Of course, the answer is the same as the example with the heads and tails: 15,504.

This approach—enumerating possible combinations from our samples—will be important when we use maximum likelihood estimation to estimate survival rates. But, for now, we can be happy that we are able to use the binomial coefficient to think about samples.

MLE: a fish tagging example

Let's illustrate the concept of maximum likelihood estimation with an example. Imagine that biologists are conducting a tag loss study of fish—before embarking on a larger study, the scientists want to know if their fish tags will stay in place or, instead, will simply fall off, which would be bad for the larger study!

So, the biologists put 10 fish in a tank, and each fish is marked with the special tag (Figure 3.2). After one day, the animals are checked. Seven of the 10 fish have retained their tags, and three of the fish have lost their tags. The biologists need to estimate the probability of tag retention over the 1-day period

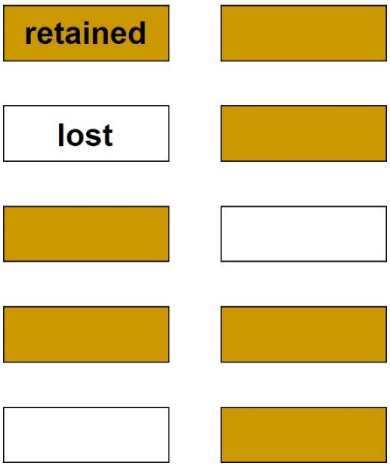

Figure 3.2: *Depiction of 10 marked fish placed in a tank. Shaded boxes represent fish that retained tags, and open boxes represent fish with lost tags.*

of time. And, they are pretty sure they should use a maximum likelihood estimation process.

However, the 12-year-old daughter of one of the scientists, Dr. Bigstuff, enters the laboratory and tells them they are thinking too hard—the answer is that the tag retention probability is 70%. All of the other scientists agree—she appears to be correct. It's obvious—you use the fraction 7/10. Seven of 10 fish remained tagged. That's 70%. Case closed, right?

Instead, Dr. Bigstuff wants to prove that his daughter is not correct. He still thinks a maximum likelihood approach would be best for this important project. He sends his daughter home to work on her mathematics homework, while he sits down to work on his own math problem.

In the above section, we learned about **binomial sampling**. We'll put that information to use here as we help Dr. Bigstuff.

> *A **binomial trial,** also known as a **Bernoulli trial**, has exactly **two outcomes**. The trial will end in either a success or a failure. When we apply this to our fish tagging, we can label the tag retention as a success and a tag loss as a failure.*

We should point out that we could also label the tag loss as the success if that made more sense (it doesn't seem to, does it?!)—but we must be clear about the definition as we move forward, and we must remember which outcome we labeled as a success.

In an experiment based on binomial trials, we know that we will have n trials and we will have y successes. And, we know that y is an integer and the value of y can be stated as: $0 \leq y \leq n$.

Now that we have labeled our success (tag retention), we can identify a parameter of interest. In binomial trials, we usually use p to represent the probability of success. And, conversely, q is the probability of failure. If there are only two outcomes of our trial, then we know that the following is true: $q = 1\text{-}p$

And, because p is a probability, p will be continuous with the value of: $0 \leq p \leq 1$

Last, because we have data, we can estimate p with a maximum likelihood estimator. Maybe Dr. Bigstuff is correct…?

Writing the likelihood statement

A binomial trial is a very simple experiment. And, as a result the "likelihood statement", or "probability statement", is also relatively simple. The general Bernoulli likelihood formula is written as follows:

$$L(y \mid N, p) = \binom{N}{y} p^{y} (1-p)^{(N-y)}$$

Formulas are nice, but it's good to be able to translate them into words—so, let's do that. On the left side of the equation, above, is the statement, $L(y|N,p)$. This can be read, "the likelihood (or probability) of observing y (the number of successes), given N (the number of

trials) and the presence of a parameter, p, equals…" Another way to say this is, "We'll be estimating the value for a parameter given some data and underlying assumptions about a simple model." That should sound like our definition for maximum likelihood estimation…

On the right side of the equation, we have a more complex statement. You should recognize the binomial coefficient:

$$\binom{N}{y}$$

We use the binomial coefficient to calculate how many ways (combinations) you could get what you are proposing (in our case: *how many ways could 7 tags be retained on 10 fish?*).

The product

$$p^{y}(1-p)^{(N-y)}$$

is the probability statement that will estimate the probability of getting our exact result (seven successes and three failures in our fish tagging experiment) in any order (could be Fish #1-7 retain tags, or it could be Fish #2-8 retain tags, etc.). If p is the probability of one success, we need to raise p to the yth power to get the probability of y successes. And, if q=1-p (representing the probability of a failure), we need to raise 1-p to the N-y power to get the probability of N-y failures.

In our fish tagging example:
y = 7 successes
N-y = 3 failures

And, for our specific experimental results with 10 trials and y observed successes, Dr. Bigstuff could arrange the equation (here we use $L(\theta|y=7)$ to represent $L(y|N,p)$ and to designate the specific results of our experiment: y=7 successes) as:

$$L(\theta \mid y=7)=\binom{10}{7}p^{7}(1-p)^{10-7}$$

We now have our **likelihood statement** and this is the first step along the path to use a maximum likelihood estimator. Remember that we said the MLE process would be logical, because we are using probabilities? And, our fish tagging example shows the relatively simple pieces that go together to construct a likelihood statement.

Maximum likelihood cookbook

Luckily for most of us, we will never have to do MLE's by hand. That is why we paid a lot of money for our computers, right?!

But, for the sake of understanding what is going on when your computer obtains a MLE, there are two basic steps in the MLE process:

- The **first step** is to **state the structure of our model and write out the likelihood function** for our experiment. We've done that for the fish tagging experiment. We had 10 trials, and it was a binomial situation with a success or a failure in each trial. And, we

wanted to estimate a parameter—the probability of success in the trial (p). Done. On to step two.

- The **second step** involves *calculus*. *Please don't close the book*. It's not that painful. To get the maximum likelihood estimate, we must **maximize the likelihood function**. Whoops! Did you notice that? Probably not—you were worried about the calculus. But, there was a word that we were looking for: ***maximum***. By maximizing the likelihood function, we will be able to determine what value of the parameter is most likely, given our data. It is all coming together—really!

> **Calculus** is a mathematical approach to study either slopes and curves (differential calculus) or areas under curves (integral calculus). In this chapter, we will use basic concepts of differential calculus. In Chapter 19, we will use concepts of integral calculus.

Maximizing the likelihood: with graphics

Let's go back to our example of 7 fish that retained their tags in an experiment of 10 tagged fish. Again, we start with this likelihood, for our 10 trials and 7 successes:

$$L(\theta \mid y = 7) = \binom{10}{7} p^7 (1-p)^{10-7}$$

If we simplify the right side of the likelihood function, we get the following (worked out in steps here):

$$L(\theta \mid y = 7) = \left[\frac{10!}{7!(10-7)!} \right] p^7 (1-p)^3$$

We can simplify this by writing out the factorial statements. If 10! = 10*9*8*7*6*5*4*3*2*1, and if 7!=7*6*5*4*3*2*1, then 10!/7! in the above equation can be simplified to 10*9*8. Then:

$$L(\theta \mid y = 7) = \left[\frac{10*9*8}{3*2*1} \right] p^7 (1-p)^3$$

And finally:

$$L(\theta \mid y = 7) = 120 p^7 (1-p)^3$$

Our "cookbook" directions above tell us that we now need to take our likelihood function and "maximize" it. There is a reason that calculus is needed at this step—it can help us find the maximum of the function that we've written.

Before we go further, we should point out that a function is simply the relationship between two things—in our case the function describes how the value of the likelihood (L) changes as p

changes in value. We can simplify the left side of the equation to help us visualize this equation better:

$$L = 120\,p^{7}(1-p)^{3}$$

And, we can graph that function by replacing p with various values between 0 and 1 (the range of possible p's (Figure 3.3).

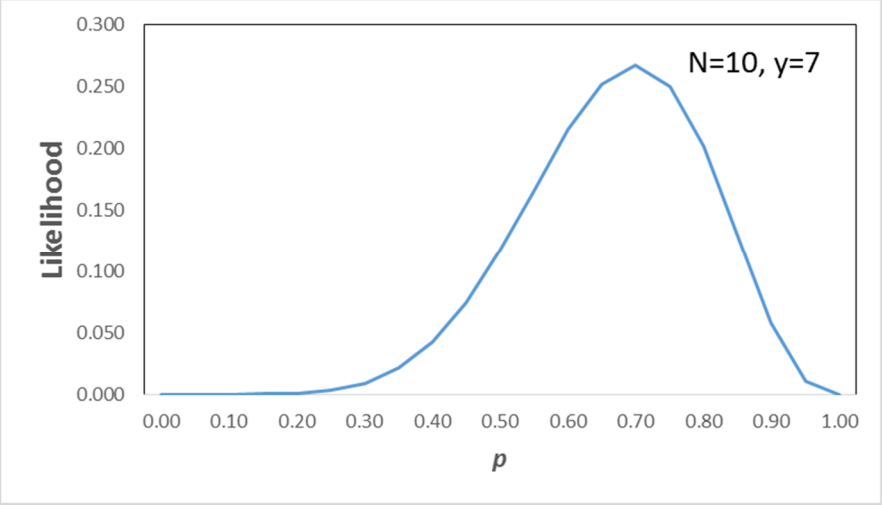

Figure 3.3: *The value of the likelihood function, $L=120p^{7}(1-p)^{3}$, for a range of values of p. The function reaches its maxima (peak) at p=0.7.*

You can try this in a spreadsheet yourself. In a blank spreadsheet, make one column labeled "p" and a second column labeled "Likelihood".

Under the "p" heading, put a column of numbers like follows:

0
0.05
0.10
…
0.95
1.00

Then, in the likelihood column, create an equation to calculate the likelihood. If you are using Microsoft Excel, the equation (in Excel format) would look something like this for the top cell (with p = 0 in the cell to its left).

= 120*(A2^7)*(1-A2)^3

If you copy that formula to the rest of the cells below in the B column, you should see values that match the ones to the right. And, you can use your Excel graphing skills to see if you can replicate Figure 3.3!

p	Likelihood
0	0
0.05	8.03789E-08
0.1	8.748E-06
0.15	0.000125915
0.2	0.000786432
0.25	0.003089905
0.3	0.009001692
0.35	0.021203015
0.4	0.042467328
0.45	0.074603106
0.5	0.1171875
0.55	0.166478293
0.6	0.214990848
0.65	0.252219625
0.7	0.266827932
0.75	0.250282288
0.8	0.201326592
0.85	0.129833721
0.9	0.057395628
0.95	0.010475059
1	0

Now that we've graphed it, can you see where the maximum of the function is? What value of p makes L (the likelihood) largest? *Interestingly, it looks like it is somewhere near p=0.7.* Perhaps Dr. Bigstuff should start listening to his daughter?!

Well, Dr. Bigstuff might claim that we can't tell exactly where L is the highest—perhaps it is at p=0.68 or p=0.72? We only used values that differed by 0.05 in our graphical approach. What is the exact maximum?

Maximizing the likelihood: with calculus

To find the exact maximum of our likelihood function, we must turn to calculus. We know you've repressed everything you learned in that class and you may not remember much…but, do you remember that if you take the **derivative** of a function, it tells you the slope at a given point along the function? Maybe you don't remember that, but take our word for it. It's true.

Using that concept, we can find where the steepest part of the curve is (that part of the curve would have the largest value for the slope), and we can find where the slope is not very steep. For example, if we draw a tangential line to the function where p=0.6, what is the slope?

That would be line A in Figure 3.4, and we can see the slope is positive and very steep.

Next, we can ask ourselves—*what would be the value of the slope at the maximum point (peak) of the curve?*

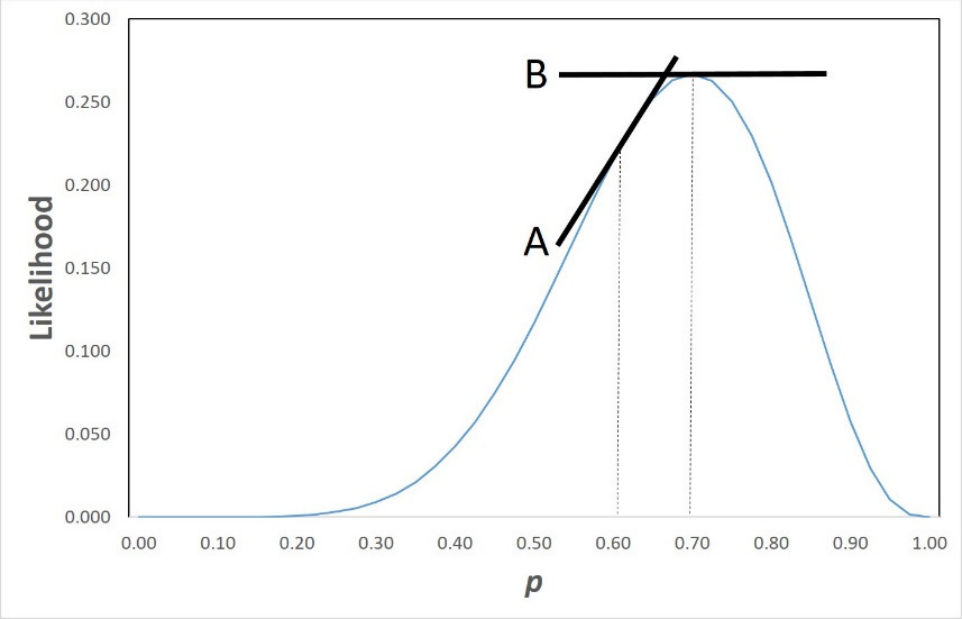

Figure 3.4: *The likelihood function shown in Figure 3.2 with depictions (A) of the slope of the function at a given value of p (here, 0.6) and (B) the slope of the function at the maxima (peak) of the function (p=0.7). At the peak, the slope=0 (no rise; flat line).*

Aha! The correct answer is—***the slope would be zero***! We can see this by looking at line B in Figure 3.4. The slope at the tip of the curve is completely flat.

Importantly, we can also notice that the maximum is the ONLY point on the curve with a slope of zero. That's important, as well.

Thus, we can use logic and put together a powerful idea—if the derivative of the likelihood function gives us its slope, and if the slope at the maximum point of the likelihood function is zero, then we should be able to set the derivative of the likelihood function equal to zero and solve for the value of *p* at which the maximum occurs. This is the crux of maximum likelihood estimation!

We can now celebrate—we've used calculus for something useful! Let's apply this to our fish tagging example. Note that we identify the derivative of a function (e.g., Z) as Z'. So, we can write our likelihood again, as before:

$$L(\theta \mid y = 7) = 120 p^7 (1-p)^3$$

Then, we indicate that we need the derivative of the function, and we set it equal to 0.

$$\left(120 p^7 (1-p)^3\right)' = 0$$

Now, we need to calculate the derivative. If you have retained a lot from calculus class, you may be able to figure this out. If not, we suggest you can quickly find the answer by googling "on-line derivative calculator" and using the on-line tool to get your answer, which is as follows (trust us, or do it on-line to see if we're correct!):

$$840(1-p)^3 \cdot p^6 - 360(1-p)^2 \cdot p^7 = 0$$

We can simplify that result as follows:

$$\frac{840}{360}(1-p) = p$$

And, then:

$$2.333 - 2.333p = p$$

And, further:

$$2.333 = 3.333p$$

Now, if we solve for *p*, we find that **the value of *p* that maximizes the likelihood function is 0.7**. Thus:

$$\hat{p} = 0.7$$

Thus, Dr. Bigstuff's daughter was correct. Her simple fraction of $p = 7/10$ or $p = y/N$ was truly a maximum likelihood estimator! See, you have been using maximum likelihood estimation since elementary school, and you didn't know it.

> **A little hat?**
>
> Throughout this primer, you will see parameters, such as p, indicated as p or as $\hat{p}$. The "hat" symbol indicates that this is an **estimate** of p rather than a theoretical value or a known value of p.

Coin toss example

Let's do another example.

In the section above, we worked with an example of 20 tosses of a coin. And, we asked about the probability of getting 5 heads and 15 tails in such a trial. This is very similar to our experiment that we've just completed with the fish and the tags! So, let's see if we can apply our knowledge to this new question.

Our coin flipping scenario is similar to the fish tagging experiment—we have a certain number of trials (20 coin flips) and we want to have 5 successes (flipping a "head", Figure 3.5). So, we can use the same likelihood formula for a Bernoulli trial:

$$L(5 \mid 20, p) = \binom{20}{5} p^5 (1-p)^{(20-5)}$$

We're using this example as we know that the probability of flipping a "head" with a "fair coin" (not weighted in any way) is 50% or p=0.50. If we replace the p's in the formula with 0.5, we find that the last portion of the equation

$$p^5 (1 - p)^{(20-5)} = 0.00000095$$

This is the probability of getting *one* of *any* combination of 5 heads and 15 tails in 20 flips, without worrying about the order.

That is a really small number (0.00000095), but we also know that we need to account for all of the combinations of ways that we could get 5 heads and 15 tails. So, we look to the first portion of the equation. As earlier in this chapter, we find:

$$\binom{20}{5} = 15504$$

Thus, there are 15504 ways to have 5 heads and 15 tails in 20 flips of a coin. And, each has a small chance of occurring (*0.00000095, to be exact!*). So, we need to multiply the two portions of this equation together—the total probability of getting some combination of 5 heads and 15 tails is equal to the number of combinations possible, multiplied by the probability of getting any one combination of 5 heads and 15 tails.

Figure 3.5: *A specific result of trials of 20 tosses of a coin referenced in the text: 5 heads and 15 tails.*

We rely on the formula to tell us that **if p = 0.5, the likelihood of getting exactly 5 heads in 20 coin flips is 0.0148.** Or, to say it another way, if we performed many, many trials of 20 coin flips, we'd expect to get 5 heads in 1.48% of our trials. And, this makes intuitive sense to us. Most of the time, we would probably get about 10 heads and 10 tails, because our coin is not defective (we assume, if p=0.5). So, getting only 5 heads should be a rare occurrence.

Proving why we need both portions of the likelihood:

Pretend for the moment that our coin flip example is simpler: we have 3 flips of a coin and we want to know the probability, or likelihood, of getting 2 heads and 1 tail.

Our likelihood expression would be: $L = \binom{3}{2} p^2 (1-p)^{(3-2)}$

The first part of the equation tells us that there are 3*2*1/2*1=3 combinations that can give us 2 heads: HHT HTH and THH.

The second part of the equation tells us that the probability of getting one of ANY combination of coin flips with 2 heads and 1 tail is $0.5^2(0.5)^1 = 0.125$.

But, we can get 2 heads and 1 tail by 3 different combinations, so we multiply the two portions together, and we find: 3*0.125=0.375. So, we have a 37.5% chance of getting 2 heads and 1 tail in 3 flips of a coin.

Because this is a small trial, we can test our answer by writing out all possible results of 3 coin flips:

HHH—*only 1 way to get 3 heads*
HHT—*three ways to get 2 heads*
HTH
THH
HTT—*three ways to get 1 head*
TTH
THT
TTT—*only 1 way to get 0 heads*

There are 8 possible results when flipping a coin 3 times. And, only three of them give us 2 heads! By simple math, 3/8 = 0.375, or 37.5% of the results. It works.

Meanwhile, back in the real world

We've just explored the use of a likelihood statement for which we know the value of p—because we know that a 'fair coin' should give us a head 50% of the time, on average.

But, we don't know the value of p when doing a study in nature. Let's change the "story" without changing the numbers—let's pretend that we release 20 marked animals and only 5 survive (Figure 3.6).

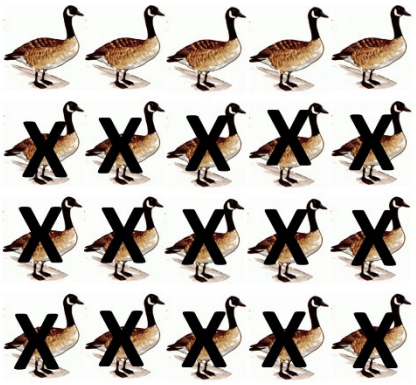

Figure 3.6: *A sample of 20 tagged animals in which only 5 survived. Dead animals are marked with an "X".*

What is p, the probability of success (or the probability of survival)?

Of course, we can use MLE to tell us the value of p that would maximize the chances of our observation occurring. That is, *given our observation of 5 animals surviving from an initial cohort of 20, what value of p is most likely?*

We still have the same formula for the likelihood:

$$L(5 \mid 20, p) = \binom{20}{5} p^5 (1-p)^{(20-5)}$$

It may help us to graphically view our function, over a range of p's and see if we can identify the value of p that maximizes the function. If we do this in a spreadsheet such as Microsoft Excel, we can see that there certainly is a peak—a maxima—to the function that we plot. What is the value of p at the maximum point? Yes, it appears to be at or near p=0.25 (Figure 3.7).

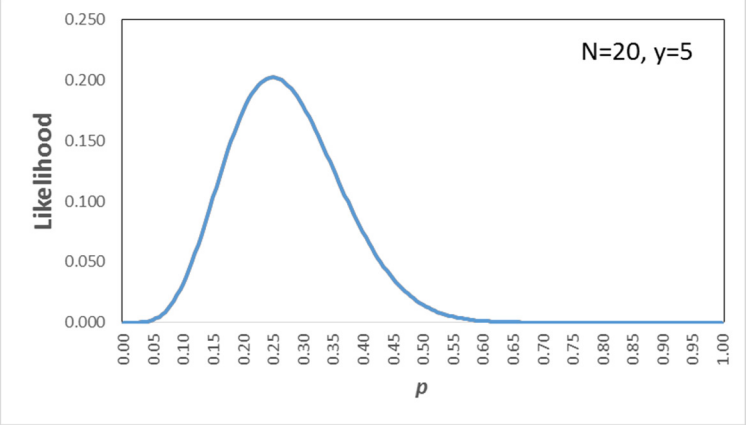

Figure 3.7: *The value of the likelihood function for the study and sample results depicted in Figure 3.6.*

In our previous fish tagging example, we looked at a graphic to see where the likelihood function appeared to have a maxima. Then, we took the likelihood equation with our specific values (7 tags retained out of 10 fish), and we worked out the value for p. We could do the same here for our 20 animals released and 5 survivors.

But, we might pause and ask ourselves—*are we going to have to do this set of calculations for every single observation in every experiment we ever use?* We hope not! It would be nice to have a general formula—an **estimator**—that would give us our estimate.

Luckily, statisticians have done some preparatory work for us. They have provided an elegant way to obtain a maximum likelihood estimator for a simple Bernoulli experiment with *N* trials and *y* successes.

If we take our general Bernoulli likelihood statement,

$$L(y \mid N, p) = \binom{N}{y} p^{y} (1-p)^{(N-y)}$$

and then take natural logarithms of the entire likelihood, we obtain the "log likelihood".

Mathematicians suggest doing the log transformation to the equation because it simplifies the analytical process—we remove the exponents when we take the log of each side of the equation to get:

$$\ln L(p \mid data) = y \ln p + (N-y) \ln(1-p)$$

Then, we take the derivative of the equation, with respect to *p*. Again, this provides us with an equation that can give us the slope of the function at any value of *p*. We are most interested in the value of *p* where the slope is zero (the maxima!), so we set the derivative equal to 0:

$$\frac{\delta[\ln L(p \mid data)]}{\delta p} = \frac{y}{p} - \frac{(N-y)}{(1-p)}$$

$$\frac{y}{p} - \frac{(N-y)}{(1-p)} = 0$$

Then, we simplify the equation to give us our maximum likelihood estimator for *p*:

$$\hat{p} = \frac{y}{N}$$

Now, if we go back to our 20 animals with 5 survivors, we find that $\hat{p}$ = 5/20, or 0.25. That matches our best guess from looking at Figure 3.6, correct?

These are fairly simple examples of maximum likelihood estimators, using a simple Bernoulli trial. Although it is simple, this is actually the estimator for known-fate survival that can be obtained from radio-marked animals! Scientists release a certain number of animals and have a certain number of survivors every day somewhere out in the field. So, even though we started with a simple example, it is useful, and you will see this concept again in later chapters.

Variance estimation, MLE-style

Before we leave the topic of maximum likelihood estimation (although we might ask in a metaphysical sense, do we ever really leave the topic of maximum likelihood estimation?!), we need to talk about variance estimation. This is important—there is no journal on earth (well perhaps a *slight* exaggeration) that will allow you to publish a parameter estimate without an estimate for the variance, the standard error, and/or a 95% confidence interval!

Luckily, if we look back at the benefits of using maximum likelihood estimation methods, we see that one advantage is that it is straightforward to estimate the variance for a parameter using MLE.

Let's start by investigating this graphically (Figure 3.8):

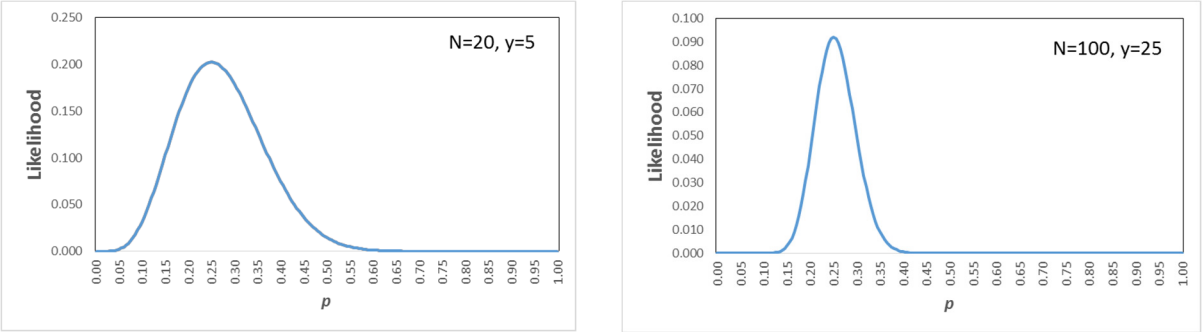

Figure 3.8: *Comparison of two similar likelihood functions, both with maxima at p=0.25. The function at left represents a study with a sample size of 20, while the function at right represents a study of 100 tagged animals.*

You'll notice that both figures appear to have maxima at p=0.25 (Figure 3.8). For reference, the figure at left is for the example we just covered—a simple Bernoulli trial of 20 animals tagged with 5 survivors and survival was estimated as $\hat{p}$ =0.25.

The difference in the curves in the two graphs (Figure 3.8) is that the one on the right is from a larger experiment. This scientist had more money for radio-tags! They released 100 animals and 25 survived. If we put those numbers into our estimator for survival, we find that again, survival was estimated as 0.25.

So, the same model structure (a Bernoulli trial) and the same estimate of the parameter. The only thing that differs is the sample size, or the number of trials (N).

Conceptually, **sampling variance** is related to the curvature of the likelihood function at is maximum. At the peak, or the maxima, is the curvature narrow or wide? It is wider on the figure to the left and narrower on the figure to the right.

We know from elementary statistics that sample size (N, the number of trials) affects the sample variance of the estimate, $\hat{p}$. We see this graphically in Figure 3.8—clearly, the sample size changes the relative likelihood of the function at various values of p. A narrower range of values for p have high likelihood values (and thus, a much narrower range of values of p are likely, given the observed data and the model structure).

Mathematically, we have to go back to calculus for the variance estimation. We've already taken the derivative of the likelihood and set it to zero to find the value of p that maximizes the likelihood function. We could call that derivative the **first derivative** because it is …*well*... the first derivative that we calculated for that likelihood.

We could also take the derivative of the first derivative (let's see…perhaps we should call it the **second derivative**?!). In calculus, the first derivative gives us information about the slope of the line (steep or shallow or flat), and the second derivative gives us information about the curvature of the line. Based on what we described above about the curvature at the maxima for the likelihood, it sounds like the second derivative of our likelihood function might be useful!

And, it is.

Specifically, if we trust the estimable Sir Ronald Fisher to guide us, we are told that the negative inverse of the second derivative provides the maximum likelihood estimate for the variance, and for our simple Bernoulli trial, the variance estimate is:

$$\text{var}(\hat{p}) = \frac{p(1-p)}{N}$$

You try it

You can have some fun with this simple maximum likelihood estimator by placing 100 beans of 2 colors (any combination) in a cup. Mix a known number of dark and light beans (any seed pods or other small colored objects will work). So, in this example you know the truth about p, the proportion of light beans (a success).

Now, take a series of samples—first take 10 beans at random from the cup. $N = 100$ and $y = 10$ (the number of light-colored beans you get in your sample). What is your estimate for p? Is it close to what you know to be the truth?

Now, use the following formula to estimate the confidence interval for your p-hat:

Lower bound of 95% confidence interval: $\hat{p} - 1.96 \cdot \sqrt{\frac{p(1-p)}{N}}$

Upper bound of 95% confidence interval: $\hat{p} + 1.96 \cdot \sqrt{\frac{p(1-p)}{N}}$

Does the 95% CI contain the true value of p? Perhaps you need a larger sample to reduce bias and increase precision? So, put the 10 beans back in the cup with the other 90 beans and mix them up. Now, remove 25 beans and estimate $\hat{p}$ and calculate the 95% CI again. Last, try removing 50 beans as your sample. N will always equal 100, but y will change: 10 to 25 to 50.

What happens to your confidence interval? It should get smaller as your sample size increases.

Conclusion

Maximum likelihood estimates will be used 'behind the scene' in the methods described throughout this primer. A basic knowledge of the process used to establish a maximum likelihood estimator is valuable for a biologist to have. If nothing else, remember that a maximum likelihood estimator is a method to find the most likely value for a parameter, given the model structure and our observed data.

To find the maximum likelihood estimate, two steps are necessary: (1) you must write the statement of likelihood for your experiment (our examples were Bernoulli trials), and (2), you must use the likelihood function and a small (*really!*) amount of calculus to find the maxima of the likelihood function.

For more information on topics in this chapter

Conroy, M. J., and J. P. Carroll. 2009. Quantitative Conservation of Vertebrates. Wiley-Blackwell: Sussex, UK.

Cooch, E., and G. White. 2014. Chapter 1: First Steps. *In* Program MARK: a gentle introduction, 12th edition, Cooch, E. and G. White, eds. Online: http://www.phidot.org/software/mark/docs/book/pdf/chap1.pdf

Donovan, T. M. and J. Hines. 2007. Exercises in occupancy modeling and estimation. Online: http://www.uvm.edu/envnr/vtcfwru/spreadsheets/occupancy.htm

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic Press, San Diego.

Citing this primer

Powell, L. A., and G. A. Gale. 2015. Estimation of Parameters for Animal Populations: a primer for the rest of us. Caught Napping Publications: Lincoln, NE.

A northern crested caracara (Caracara cheriway) is tagged with an aluminum leg band in Florida, USA. Photo by Jennifer Smith, used with permission.