CHAPTER 1

First steps. . .

We'll introduce the basic idea for analysis of data from marked individuals by means of a simple example. Suppose you are interested in exploring the costs of reproduction on survival of some species of your favorite taxa (say, a species of bird). The basic idea is pretty simple: an individual that spends a greater proportion of available energy on breeding may have less available for other activities which may be important for survival. In this case, individuals putting more effort into breeding (i.e., producing more offspring) may have lower survival than individuals putting less effort into breeding. On the other hand, it might be that individuals that are of better 'quality' are able to produce more offspring, such that there is no relationship between 'effort' and survival. You decide to reduce the confounding effects of the 'quality' hypothesis by doing an experiment. You take a sample of individuals who all produce the same number of offspring (the idea being, perhaps, that if they had the same number of offspring in a particular breeding attempt, that they are likely to be of similar quality). For some of these individuals, you increase their 'effort' by adding some offspring to the nest (i.e., more mouths to feed, more effort expended feeding them). For others, you reduce effort by removing some offspring from the nest (i.e., fewer mouths to feed, less effort spent feeding them). Finally, for some individuals, you do not change the number of offspring, thus creating a control group.

As described, you've set up an 'experiment', consisting of a control group (unmanipulated nests), and 2 treatment groups: one where the number of offspring has been reduced, and one where the number of offspring has been increased. For convenience, call the group where the number of offspring was increased the 'addition' group, and call the group where the number of offspring was reduced the 'subtraction' group. Your hypothesis might be that the survival probability of the females in the 'addition' group should be lower than the control (since the females with enlarged broods might have to work harder, potentially at the expense of survival), whereas the survival probability of the females in the 'subtraction' group should be higher than the control group (since the females with reduced broods might not have to work as hard as the control group, potentially increasing their survival). To test this hypothesis, you want to estimate the survival of the females in each of the 3 groups. To do this, you capture and individually mark the adult females at each nest included in each of the treatment groups (control, additions, subtractions). You release them, and come back at some time in future to see how many of these marked individuals are 'alive' (the word 'alive' is written parenthetically for a reason which will be obvious in a moment).

Suppose you mark 50 individuals in each of the 3 groups. Call this time (t). Then, at some later time (call it t+1), you go back out in the field and encounter alive 30 of the marked individuals from the 'additions' treatment, 35 of the marked individuals from the control group, and 30 individuals from the 'subtractions' treatment. The 'encounter data' are tabulated at the top of the next page.

group	t	$t+1$
additions	50	30
control	50	38
subtractions	50	30

Hmm. This seems strange. While you predicted that the 2 treatment groups would differ from the controls, you did not predict that the results from the two treatments would be the same. What do these results indicate? Well, of course, you could resort to the time-honored tradition of trying to concoct a parsimonious ‘post-hoc adaptationist’ story to try to demonstrate that (in fact) these results ‘made perfect sense’, according to some ‘new twist to underlying theory’. However, there is another possibility - namely, that the analysis has not been thoroughly understood, and as such, interpretation of the results collected so far needs to be approached very cautiously.

1.1. Return ‘rates’

Let’s step back for a moment and think carefully about our experiment - particularly, the analysis of ‘survival’. In our study, we marked a sample of individual females, and simply counted the numbers of those females that were subsequently seen again on the next sampling occasion. The implicit assumption is that by comparing relative proportions of ‘survivors’ in our samples (perhaps using a simple χ^2 test), we will be testing for differences in ‘survival probability’. However (and this is the key step), is this a valid assumption? Our data consist of the number of marked and released individuals that were encountered again at the second sampling occasion. While it is obvious that in order to be seen on the second occasion, the marked individual must have survived, is there anything else that must happen? The answer (perhaps obviously, but in case it isn’t) is ‘yes’ - the number of individuals encountered on the second sampling occasion is a function of 2 probabilities: the probability of survival, and the probability that conditional on surviving, that the surviving individual is encountered. While the first of these 2 probabilities is obvious (and is in fact what we’re interested in), the second may not be. This second probability (which we refer to generically as the ‘encounter probability’) is the probability that given that the individual is alive and in the sample, that it is in fact encountered (e.g., seen, or ‘visually encountered’). In other words, simply because an individual is alive and in the sampling area may not guarantee that it is encountered. So, the proportion of individuals that were encountered alive on the second sampling occasion (which is often referred to in the literature as ‘return rate’[*]) is the product of 2 different probability processes: the probability of surviving and returning to the sampling area (which we’ll call ‘apparent’ or ‘local’ survival), and the probability of being encountered, conditional on being alive an in the sample (which we’ll call ‘encounter probability’). So, ‘return rate’ = ‘survival probability’ × ‘encounter probability’. Let’s let ϕ (pronounced ‘*fee*’ or ‘*fie*’, depending on where you come from) represent the ‘local survival probability’, and p represent the ‘encounter probability’. Thus, we would write ‘return rate’ = ϕp.

So, why do we care? We care because this complicates the interpretation of ‘return rates’ - in our example, differences in ‘return rates’ could reflect differences in the probability of survival, or they could reflect differences in encounter probability, or both! Similarly, lack of differences in ‘return rates’ (as we see when comparing the ‘additions’ and ‘subtractions’ treatment groups in our example) may not indicate ‘no differences in survival’ (as one interpretation) – there may in fact be differences in survival, but corresponding differences in encounter probability, such that their products (‘return

[*] The term ‘return rate’ is something of a misnomer, since it is not a *rate*, but rather a *proportion*. However, because the term ‘return rate’ is in wide use in the literature, we will continue to use it here.

rate') are equal. For example, in our example study, the 'return rate' for both the 'additions' and 'subtractions' treatment groups is the same: (30/50) = 0.6. Our initial 'reaction' might have been that these data did not support our hypothesis predicting difference in survival between the 2 groups. However, suppose that in fact the 'treatment' (i.e., manipulating the number of offspring in the nest) not only influenced survival probability (as was our original hypothesis), but also potentially influenced encounter probabilities? For example, suppose the true survival probability of the 'additions' group was $\phi_{add} = 0.65$ (i.e., a 65% chance of surviving from t to t+1), while for the 'subtractions' group, the survival probability is $\phi_{sub} = 0.80$ (i.e., an 80% chance of surviving). However, in addition, suppose that the encounter probability for the 'additions' group was $p_{add} = 0.923$ (i.e., a 92.3% chance that a marked individual will be encountered, conditional on it being alive and in the sampling area), while for the 'subtractions' group, the encounter probability was $p_{sub} = 0.75$ (we'll leave it to proponents of the adaptationist paradigm to come up with a 'plausible' explanation for such differences). While there are clear differences between the 2 groups, the products of the 2 probabilities are the same: $(0.65 \times 0.923) = 0.6$, and $(0.8 \times 0.75) = 0.6$. In other words, it is difficult to compare 'return rates', since differences (or lack thereof) could reflect differences or similarities in the 2 underlying probabilities (survival probability, and encounter probability).

1.2. A more robust approach

How do we solve this dilemma? Well, the solution we're going to focus on here (and essentially for the next 800 pages or so) is to collect more data, and using these data, separately estimate all of the probabilities (at least, when possible) underlying the encounters of marked individuals. Suppose for example, we collected more data for our experiment, on a third sampling occasion (at time $t + 2$). On the third occasion, we encounter individuals marked on the first occasion. But, perhaps some of those individuals encountered on the third occasion were not encountered on the second occasion. How would we be able to use these data? First, we introduce a simple bookkeeping device, to help us keep track of our 'encounter' data (in fact, we will use this bookkeeping system throughout the rest of the book – discussed in much more detail in Chapter 2). We will 'keep track' of our data using what we call '*encounter histories*'. Let a '`1`' represent an encounter with a marked individual (in this example, we're focusing only on 'live encounters'), and let a '`0`' indicate that a particular marked individual was not seen on a particular occasion. Now, recall from our previous discussion that a '`0`' could indicate that the individual had in fact died, but it could also indicate that the individual was in fact still alive, but simply not encountered (the problem we face, as discussed, is how to differentiate between the two possibilities). For our 3 occasion study, where individuals were uniquely marked on the first occasion only, there are 4 possible encounter histories:

encounter history	*interpretation*
`111`	captured and marked on the first occasion, alive and encountered on the second occasion, alive and encountered on the third occasion
`110`	captured and marked on the first occasion, alive and encountered on the second occasion, and either (i) dead by the third occasion, or (ii) alive on the third occasion, but not encountered
`101`	captured and marked on the first occasion, alive and not encountered on the second occasion, and alive and encountered on the third occasion

100	captured and marked on the first occasion, and either (i) dead by the second occasion, (ii) alive on the second occasion, and not encountered, and alive on the third occasion and not encountered, (iii) alive on the second occasion, and not encountered, and dead by the third occasion

You might be puzzled by the verbal explanation of the third encounter history - 101. How do we know that the individual is alive at the second occasion, if we didn't see it? Easy - we come to this conclusion logically, since we saw it alive at the third occasion. And, if it was alive at occasion 3, then it must also have been alive at occasion 2. But, we didn't see it on occasion 2, even though we know (logically) that it was alive. This, in fact, is one of the key pieces of logic – the individual was alive at the second occasion but not seen. If p is the probability of detecting (encountering) an individual given that it is alive and in the sample, then $(1-p)$ is the probability of missing it (i.e., not detecting it). And clearly, for encounter history '101', we 'missed' the individual at the second occasion.

All we need to do next is take this basic idea, and formalize it. As written (above), you might see that each of these encounter histories could occur due to a specific sequence of events, each of which has a corresponding probability. Let ϕ_i be the probability of surviving from time (i) to (i+1), and let p_i be the probability of encounter at time (i). Again, if p_i is the probability of encounter at time (i), then $(1-p_i)$ is the probability of not encountering the individual at time (i).

Thus, we can re-write the preceding table as

encounter history	*probability of encounter history*
111	$\phi_1 p_2 \phi_2 p_3$
110	$\phi_1 p_2 \left[\phi_2(1-p_3)+(1-\phi_2)\right]$ $=\phi_1 p_2(1-\phi_2 p_3)$
101	$\phi_1\,(1-p_2)\,\phi_2 p_3$
100	$(1-\phi_1)+\phi_1(1-p_2)(1-\phi_2)+\phi_1(1-p_2)\phi_2(1-p_3)$ $=1-\phi_1 p_2-\phi_1(1-p_2)\phi_2 p_3$

(If you don't immediately see how to derive the probability expressions corresponding to each encounter history, not to worry: we will cover the derivations in much more detail in later chapters).

So, for each of our 3 treatment groups, we simply count the number of individuals with a given encounter history. Then what? Once we have the number of individuals with a given encounter history, we use these frequencies to *estimate* the probabilities which give rise to the observed frequency. For example, suppose for the 'additions' group we had $N^{111}=7$ (where N^{111} is the number of individuals in our sample with an encounter history of '111'), $N^{110}=2$, $N^{101}=5$, and $N^{100}=36$. So, of the 50 individuals marked at occasion 1, only $(7+2+5)=14$ individuals were subsequently encountered alive (at either sampling occasion 2, sampling occasion 3, or both), while 36 were never seen again. Suppose for the 'subtractions' group we had $N^{111}=5$, $N^{110}=7$, $N^{101}=2$, and $N^{100}=36$. Again, 14 total individuals encountered alive over the course of the study.

However, even though both treatment groups (additions and subtractions) have the same overall 3-year return rate $(14/50=0.28)$, we see clearly that the frequencies of the various encounter histories differ between the groups. This indicates that there are differences among encounter occasions in survival probability, or encounter probability (or both) between the 2 groups, despite no difference in overall return rate. The challenge, then, is how to estimate the various probabilities (parameters) in the probability expressions, and how to determine if these parameter estimates are different between the 2 treatment groups.

An *ad hoc* way of getting at this question involves comparing ratios of frequencies of different encounter ratios. For example,

$$\frac{N^{111}}{N^{101}} = \frac{\phi_1 p_2 \phi_2 p_3}{\phi_1 (1 - p_2) \phi_2 p_3} = \frac{p_2}{1 - p_2}$$

So, for the 'additions' group, $(N^{111}/N^{101}) = (7/5) = 1.4$. Thus, $\hat{p}_{(2,add)} = 0.583$. In contrast, for the 'subtractions' group, $(N^{111}/N^{101}) = (5/2) = 2.5$. Thus, $\hat{p}_{(2,sub)} = 0.714$. Once we have estimates of p_2, we can see how we could substitute these values into the various probability expressions to solve for some of the other parameter (probability) values. However, while this is reasonably straightforward (at least for this very simple example), what about the question of 'is this difference between the two different $\hat{p}_2$ values meaningful/significant?'. To get at this question, we clearly need something more - in particular we need to be able to come up with estimates of the uncertainty (variance) in our parameter estimates. To do this, we need a robust statistical tool.

1.3. Maximum likelihood theory - the basics

Fortunately, we have such a tool at our disposal. Analysis of data from marked individuals involves making inference concerning the probability structure underlying the sequence of events that we observe. Maximum likelihood (ML) estimation (courtesy of Sir Ronald Fisher) is the workhorse of analysis of such data. While it is possible to become fairly proficient at analysis of data from marked individuals without any real formal background in ML theory, in our experience at least a passing familiarity with the concepts is helpful. The remainder of this (short) introductory chapter is intended to provide a simple (very) overview of this topic. The standard 'formal' reference is the 1992 book by AWF Edwards ('*Likelihood*', Johns Hopkins University Press). Readers with significant backgrounds in the theory will want to skip this chapter, and will absolutely refrain from comment as to the necessary simplifications we make.

So here we go. . .the basics of maximum likelihood theory without (much) pain. . .

1.3.1. Why maximum likelihood?

The method of maximum likelihood provides estimators that are both reasonably intuitive (in most cases) and several have some 'nice properties' (at least statistically):

1. The method is very broadly applicable and is simple to apply.
2. Once a maximum-likelihood estimator is derived, the general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference. More technically:
    (a) maximum-likelihood estimators are *consistent*.
    (b) they are asymptotically unbiased (although they may be biased in finite samples).
    (c) they are asymptotically *efficient* - no asymptotically unbiased estimator has a smaller

asymptotic variance.

(d) they are asymptotically normally distributed - this is particularly useful since it provides the basis for a number of statistic 'tests' based on the normal distribution (discussed in more detail in Chapter 4).

(e) if there is a *sufficient statistic* for a parameter, then the maximum likelihood estimator of the parameter is a function of a sufficient statistic.*

3. A disadvantage of the maximum likelihood method is that it frequently requires strong assumptions about the structure of the data.

1.3.2. Simple estimation example - the binomial coefficient

We will introduce the basic idea behind maximum likelihood (ML) estimation using a simple, and (hopefully) familiar example: a binomial model with data from a flip of a coin. Much of the analysis of data from marked individuals involves ML estimation of the probabilities defining the occurrence of one or more events. Probability events encountered in such analyses often involve binomial or multinomial distributions. As you might appreciate, there is a simple, logical connection between binomial probabilities, and analysis of data from marked individuals, since many of the fundamental parameters we are interested in are 'binary' (having 2 possible states). For example, survival probability (live or die), detection probability (seen or not seen), and so on. Like a coin toss (head or tail), the estimation methods used in the analysis of data from marked individuals are deeply rooted in basic binomial theory. Thus, a brief review of this subject is in order.

To understand binomial probabilities, you need to first understand *binomial coefficients*. Binomial coefficients are commonly used to calculate the number of ways (combinations) a sample size of n can be taken without replacement from a population of N individuals.

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{1.1}$$

This is read as 'the number of ways (combination) a sample size of n can be taken (without replacement) from a population of size N'. Think of N as the number of organisms in a defined population, and let n be the sample size, for example. Recall that the '!' symbol means *factorial* (e.g., $5! = 5\times4\times3\times2\times1 = 120$). A quick example – how many ways can a sample of size 2 (i.e., $n = 2$) be taken from a population of size 4 (i.e., $N = 4$)? Just to confirm we're getting the right answer, let's first derive the answer by 'brute force'. Let the individuals in the sample all have unique marks: call them individuals **A**, **B**, **C** and **D**, respectively. So, given that we sample 2 at a time, without replacement, the possible combinations we could draw from the 'population' are:

AB	AC	AD	BC	BD	CD
BA	CA	DA	CB	DB	DC

So, 6 total different combinations are possibly selected (6, not 12 – the pair in each column are equivalent; e.g., AB and BA are treated as equivalent).

* Sufficiency is the property possessed by a statistic, with respect to a parameter, when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter. For example, the arithmetic mean is sufficient for the mean (μ) of a normal distribution with known variance. Once the sample mean is known, no further information about μ can be obtained from the sample itself.

So, does this match with $\binom{4}{2}$?

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{24}{2(2)} = \frac{24}{4} = 6$$

Nice when things work out, eh? OK, to continue - we use the binomial coefficient to calculate the *binomial probability*. For example, what is the probability of 5 heads in 20 tosses of a fair coin. Each individual coin flip is called a *Bernoulli trial*, and if the coin is fair, then the probability of getting a head is $p = 0.5$, while the probability of getting a tail is $(1 - p) = 0.5$ (commonly denoted as q). So, given a fair coin, and $p = q = 0.5$, then the probability of y heads in N flips of the coin is:

$$f(y \mid N, p) = \binom{N}{y} p^y (1-p)^{(N-y)} \tag{1.2}$$

The left-hand side of the equation is read as 'the probability of observing y events given that we do the experiment – toss the coin N times, and given that the probability of a head in any given experiment (i.e., toss of the coin) is p'. Given that $N = 20$, and $p = 0.5$, then the probability of getting exactly 5 heads in 20 tosses of the coin is:

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

First, we calculate $\binom{20}{5} = 15{,}504$ (note: 20! is a **huge** number). If $p = 0.5$, then $f(5 \mid 20, 0.5) = (15{,}504 \times 0.03125 \times 0.000030517578125) = 0.0148$. So, there is a 1.48% chance of having 5 heads out of 20 coin flips, if $p = 0.5$.

Now, in this example, we are assuming that we *know* both the number of times that we toss the coin, and (critically) the probability of a head in a single toss of the coin. However, if we are studying the survival of some organism, for example, what information on the left side of the probability equation (above) would we know? Well, hopefully we know the number of individuals marked (N). Would we know the survival probability (in the above, the survival probability would correspond to p - later, we'll call it S)? No! Clearly, this is what we're trying to estimate.

So, given the number of marked individuals (N) at the start of the study and the number of individuals that survive (y), how can we estimate the survival probability p? Easy enough, actually - we simply work 'backwards' (more or less). We find the value of p that maximizes the *likelihood* ($\mathcal{L}$) that we would observe the data we did.* So, for example, what would the value of p have to be to give us the observed data?

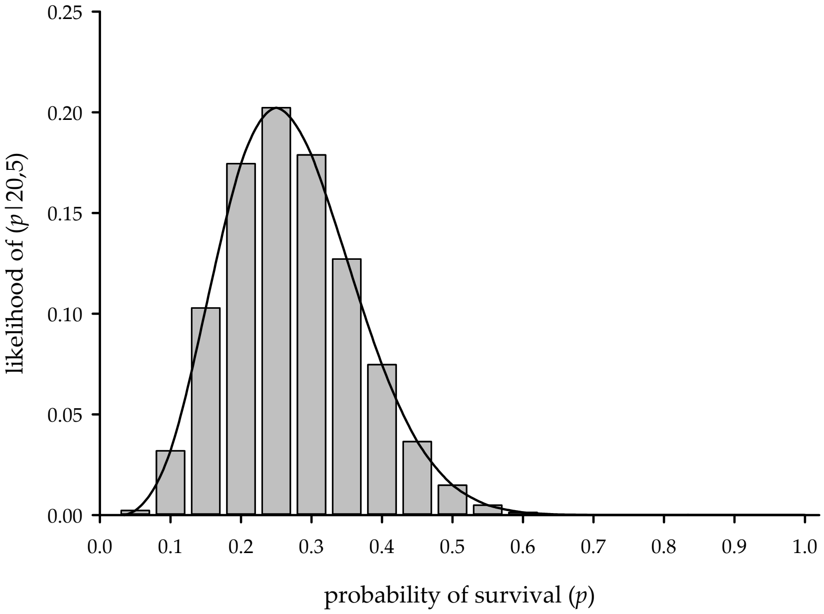
Formally, we write this as:

$$\mathcal{L}(p \mid N, y) = \binom{N}{y} p^y (1-p)^{(N-y)} \tag{1.3}$$

We notice that the right-hand side of eqn. (1.3) is identical to what it was before in eqn. (1.2) - but the left hand side is different in a subtle, but critical way. We read the left-hand side now as 'the likelihood $\mathcal{L}$ of survival probability p given that N individuals were released and that y survived'. Now, suppose $N = 20$, and that we see 5 individuals survive (i.e., $y = 5$). What would p have to be to maximize the chances of this occurring? We'll try a 'brute force' approach first, simply seeing what happens if we

* The word 'likelihood' is often used synonymously for 'probability' but in statistical usage, they are not equivalent. One may ask 'If I were to flip a fair coin 10 times, what is the *probability* of it landing heads-up every time?' or 'Given that I have flipped a coin 10 times and it has landed heads-up 10 times, what is the *likelihood* that the coin is fair?' but it would be improper to switch 'likelihood' and 'probability' in the two sentences.

set $p = 0, 0.1, 0.2, \ldots,$ and so on. Look at the following plot of the binomial probability calculated for different values of p:



As you see, the probability of 'observing 5 survivals out of 20 individuals' rises to a maximum when p is 0.25. In other words, if p, which is unknown, were 0.25, then this would correspond to the maximal probability of observing the data of 5 survivors out of 20 released individuals. This graph shows that some values of the unknown parameter p are 'relatively unlikely' (i.e., those with low likelihoods), given the data observed. The value of the parameter p at which this graph is at a maximum is the most likely value of p (the probability of a head), given the data. In other words, the chances of actually observing 11 heads and 5 tails are maximal when p is at the maximum point of the curve, and the chances are less when you move away from this point.

While graphs are useful for getting a 'look' at the likelihood, we prefer a more elegant way to estimate the parameter. If you remember any of your basic calculus at all, you might recall that what we want to do is find the maximum point of the likelihood function. Recall that for any function $y = f(x)$, we can find the maximum inflection point over a given domain by setting the first derivative dy/dx to zero and solving. This is exactly what we want to do here, except that we have one preliminary step - we 'could' take the derivative of the likelihood function as written, but it is simpler to convert everything to logarithms first. The main reason to do this is because it simplifies the analytical side of things considerably. The log-transformed likelihood, now referred to as a 'log-likelihood', is denoted as $\ln \mathcal{L}\,(q \mid \text{data})$.

Recall that our expression is

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

The binomial coefficient in this equation is a constant (i.e., does not depend on the unknown parameter p), and so we can ignore it, and express this equation in log terms as:

$$\mathcal{L}(p \mid \text{data}) = p^y(1-p)^{(N-y)} \;\rightarrow\; \ln \mathcal{L}(p \mid \text{data}) = y\ln(p) + (N-y)\ln(1-p)$$

Note that we've written the left-hand side in a sort of short-hand notation - 'the likelihood $\mathcal{L}$ of the parameter p, given the data' (which in this case consist of 5 survivors out of 20 individuals). So, now the equation we're interested in is:

$$\ln \mathcal{L}(p \mid \text{data}) = y\ln(p) + (N-y)\ln(1-p)$$

So, all you need to do is differentiate this equation with respect to the unknown parameter p, set equal to zero, and solve.

$$\frac{\partial\left[\ln \mathcal{L}(p \mid \text{data})\right]}{\partial p} = \frac{y}{p} - \frac{(N-y)}{(1-p)} = 0$$

So, solving for p, we get:

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

Thus, the value of parameter p which maximizes the likelihood of observing $y = 5$ given $N = 20$ (i.e., $\hat{p}$, the maximum likelihood estimate for p) is the same as our intuitive estimate: simply, y/N. Now, your intuition probably told you that the 'only' way you could estimate p from these data was to simply divide the number of survivors by the total number of animals. But we're sure you're relieved to learn that $5/20 = 0.25$ is also the MLE for the parameter p.

begin sidebar

closed and non-closed MLE

In the preceding example, we considered the MLE for the binomial likelihood. In that case, we could 'use algebra' to 'solve' for the parameter of interest ($\hat{p}$). When it is possible to derive an 'analytical solution' for a parameter (or set of parameters for likelihoods where there are more than one parameter), then we refer to the solution as a solution in '*closed form*'. Put another way, there is a closed form solution for the MLE for the binomial likelihood.

However, not all likelihoods have closed form solutions. Meaning, the MLE cannot be derived 'analytically' (generally, by taking the derivative of the likelihood and solving at the maximum, as we did in the binomial example). MLE's that cannot be expressed in closed form need to be solved numerically. Here is a simple example of a likelihood that cannot be put in closed form. Suppose we are interested in estimating the abundance of some population. We might intuitively understand that unless we are sure that we are encountering the entire population in our sample, then the number we encounter (the 'count' statistic; i.e., the number of individuals in our sample) is a fraction of the total population. If p is the probability of encountering any one individual in a population, and if n is the number we encounter (i.e., the number of individuals in our sample from the larger population), then we might intuitively understand that our canonical estimator for the size of the larger population is simply (n/p). For example, if there is a 50% chance of encountering an individual in a sample, and we encounter 25 individuals, then our estimate of the population size is $\hat{N} = (25/0.5) = 50$. (*Note*: we cover abundance estimation in detail in Chapter 14.)

Now, suppose you are faced with the following situation. You are sampling from a population for which you'd like to derive an estimate of abundance. We assume the population is 'closed' (no entries or exits while the population is being sampled). You go out on a number of sampling 'occasions', and capture a sample of individuals in the population. You uniquely mark each individual, and release it back into the population. At the end of the sampling, you record the total number of individuals encountered at least once – call this M_{t+1}. Now, if the canonical estimator for abundance is $\hat{N} = (n/p)$, then $\hat{p} = (n/N)$. In other words, if we knew the size of the population N then we could derive a simple estimate of the encounter probability p by dividing the number encountered in the sample n into the size of the population. Remember, p is the probability of encountering an individual. Thus, the probability of 'missing' an individual (i.e., not encountering it) is simple $(1 - p) = 1 - (n/N)$.

So, over t samples, we can write

$$\left(1 - \frac{n_1}{N}\right)\left(1 - \frac{n_2}{N}\right)\ldots\left(1 - \frac{n_t}{N}\right) = (1 - p_1)(1 - p_2)\ldots(1 - p_t)$$

where p_i is the encounter probability at time i, and n_i is the number of individuals caught at time i.

If you think about it for a moment, you'll see that the product on right-hand side is the overall probability that an individual is not caught – not even once – over the course of the study (i.e., over t total samples). Remember from above that we defined M_{t+1} as the number of individuals caught at least once. So, we can write

$$\left(1 - \frac{M_{t+1}}{N}\right) = \left(1 - \frac{n_1}{N}\right)\left(1 - \frac{n_2}{N}\right)\left(1 - \frac{n_3}{N}\right)\cdots\left(1 - \frac{n_t}{N}\right)$$

In other words, the LHS and RHS both equal the probability of never being caught – not even once. Now, if you had estimates ofp_i for each sampling occasion i, then you could write

$$\begin{aligned}\left(1 - \frac{M_{t+1}}{N}\right) &= (1 - p_1)(1 - p_2)\ldots(1 - p_t)\\ \frac{M_{t+1}}{N} &= 1 - (1 - p_1)(1 - p_2)\ldots(1 - p_t)\\ \hat{N} &= \frac{M_{t+1}}{1 - (1 - p_1)(1 - p_2)\ldots(1 - p_t)}\end{aligned}$$

So, the expression is rewritten in terms of N – analytical solution – closed form, right? Not quite. Note that we said *if* you had estimates of p_i. In fact, you don't. All you have is the count statistic (i.e., the number of individuals captured on each sampling occasion, n_i). So, in fact, 'all we have' are the count data (i.e., $M_{t+1}, n_1, n_2 \ldots n_t$), which (from above) we relate algebraically in the following:

$$\left(1 - \frac{M_{t+1}}{N}\right) = \left(1 - \frac{n_1}{N}\right)\left(1 - \frac{n_2}{N}\right)\left(1 - \frac{n_3}{N}\right)\cdots\left(1 - \frac{n_t}{N}\right)$$

It is not possible to 'solve' this equation so that only the parameter N appears on the LHS, while all the other terms (representing data – i.e., $M_{t+1}, n_1, n_2 \ldots n_t$) appear on the RHS. Thus, the estimator for N cannot be expressed in closed form.

However, the expression does have a solution – but it is a solution we must derive *numerically*, rather than *analytically*. In other words, we must use numerical, iterative methods to find the value of N that 'solves' this equation. That value of N is the MLE, and would be denoted as $\hat{N}$.

Consider the following data:

$$n_1 = 30, n_2 = 15, n_3 = 22, n_4 = n_t = 45, \text{ and } M_{t+1} = 79$$

Thus, one wants the value of N that 'solves' the equation

$$\left(1 - \frac{79}{N}\right)=\left(1 - \frac{30}{N}\right)\left(1 - \frac{15}{N}\right)\left(1 - \frac{22}{N}\right)\left(1 - \frac{45}{N}\right)$$

One could try to solve this equation by 'trial and error'. That is, one could plug in a guess for population size and see if the LHS = RHS (not very likely unless you can guess very well). Thinking about the problem a bit, one realizes that, logically, $N \geq M_{t+1}$ (i.e., the size of the population N must be at least as large as the number of unique individuals caught at least once, $M+t+1$). So, at least, one has a lower bound (in this case, 79 if we restrict the parameter space to integers). If the first guess for N does not satisfy the equation, one could try another guess and see if that either (1) satisfies the equation or (2) is closer than the first guess. The log-likelihood functions for many (but not all) problems are unimodal (for the exponential family); thus, you can usually make a new guess in the right direction.

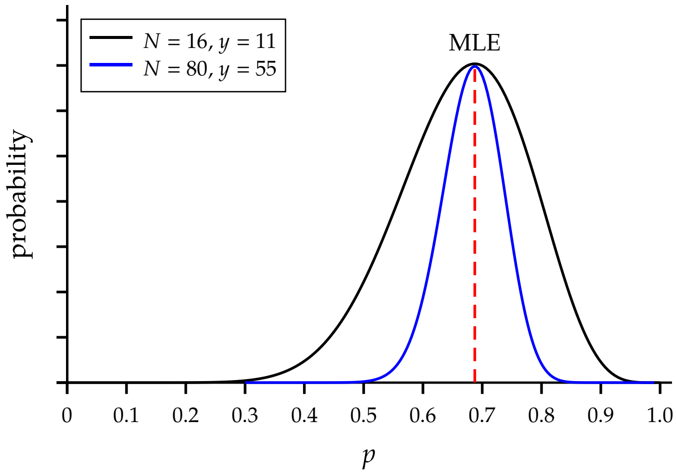
One could keep making guesses until a value of N (an integer) allows the LHS = RHS, and take this value as the MLE, $\hat{N}$. Clearly, the 'trial-and-error' method will unravel if there is more than 1 or 2 parameters. Likewise, plotting the log-likelihood function is useful only when 1 or 2 parameters are involved. We will quickly be dealing with cases where there are 30-40 parameters, thus we must rely on efficient computer routines for finding the maximum point in the multidimensional cases. Clever search algorithms have been devised for the 1-dimensional case. Computers are great at such routine computations and the MLE in this case can be found very quickly. Many (if not most) of the estimators we will work with cannot be put in closed form, and we will rely on computer software – namely, program **MARK** – to compute MLEs numerically.

end sidebar

Why go to all this trouble to derive an estimate for p? Well the maximum likelihood approach also has other uses - specifically, the ability to *estimate* the sampling variance. For example, suppose you have some data from which you have estimated that $\hat{p} = 0.6875$. Is this 'significantly different' (by some criterion) from, say, 0.5? Of course, to address this question, you need to consider the sampling variance of the estimate, since this is a measure of the uncertainty we have about our estimate. How would you do this? Of course, you might try the 'brute force' approach and simply repeat your 'experiment' a large number of times. Each time, derive the estimate of p, and then calculate a mean and variance of the parameter. While this works, there is a more elegant approach - again using ML theory and a bit more calculus (fairly straightforward stuff).

Conceptually, the sampling variance is related directly to the curvature of the likelihood at its maximum. Why? Consider the following: let's say we release 16 animals, and observe 11 survivors. What would the MLE estimate of p be? Well, we now know it is $(y/N) = (11/16) = 0.6875$. What if we had released 80 animals, instead of 16? Suppose we did this experiment, and observed 55 survivors (i.e., the expected values assuming $p = 0.6875$). What would the likelihood look like in this case? Well, clearly, the maximum of the likelihood in both 'experiments' should occur at precisely the same point: 0.6875.

But what about the 'shape' of the curve. In the following, we plot the likelihoods for both experiments ($N = 16$ and $N = 80$ respectively).



The graph shows clearly that the larger sample size ($N = 80$) results in a 'narrower' function around the ML parameter estimate $\hat{p} = 0.6875$. If the sampling variance is related to the degree of curvature of the likelihood at its maximum, then we would anticipate the sampling variance of the parameter in these 2 experiments to be quite different, given the apparent differences in the likelihood functions.

What is the basis for stating that 'variance is related to curvature'? Think of it this way - values at a given distance from the MLE are relatively 'unlikely' - the degree to which they are unlikely is a function of how rapidly the curve drops away from the maximum as you move away from the MLE (i.e., the 'steepness' of the curve on either side of the MLE).

How do we address this question of 'curvature' analytically? Well, again we can use calculus. We use the first derivative of the likelihood function to find the point on the curve where the rate of change was 0 (i.e., the maximum point on the function). This first derivative of the likelihood is known as Fisher's *score function*.

We can then use the derivative of the score function with respect to the parameter(s) (i.e., the second derivative of the likelihood function, which is known as the *Hessian*), evaluated at the estimated value of the parameter (p, in this case), to 'tell us something about the curvature' at this point. In fact, more than just the curvature, Fisher showed that the negative inverse of the second partial derivative of the log-likelihood function (i.e., the negative inverse of the Hessian), evaluated at the MLE, is the MLE of the variance of the parameter. This negative inverse of the Hessian, evaluated at the MLE, is known as the *information function, or matrix*. For our example, our estimate of the variance of p is

$$\widehat{\text{var}}(\hat{p}) = \left[-\left(\frac{\partial^2 \ln \mathcal{L}(p \mid \text{data})}{\partial p^2}\right)\right]^{-1}_{p=\hat{p}}$$

So, we first find the second derivative of the log-likelihood (i.e., the Hessian):

$$\frac{\partial^2 \mathcal{L}}{\partial p^2} = -\frac{y}{p^2} - \frac{N-y}{(1-p)^2}$$

We evaluate this second derivative at the MLE, by substituting $y = pN$ (since $\hat{p} = y/N$). This gives

$$\begin{aligned} \left.\frac{\partial^2 \mathcal{L}}{\partial p^2}\right|_{y=pN} &= -\frac{Np}{p^2} - \frac{N(1-p)}{(1-p)^2} \\ &= -\frac{N}{p(1-p)} \end{aligned}$$

The variance of p is then estimated as the negative inverse of this expression (i.e., the information function, or matrix), such that:

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

Some of you may recognize this as the often-used estimator of the variance of a binomial proportion - it is in all the 'stats books'. But now you can sleep more easily knowing how it was derived!

So, how do the sampling variances of our 2 experiments compare? Clearly, since p and (*1-p*) are the same in both cases (i.e., same ML estimate for $\hat{p}$), the only difference is in the denominator, N. Since $N = 80$ is obviously larger than $N = 16$, we know immediately that the sampling variance of the larger sample will be smaller (0.0027) than the sampling variance of the smaller sample (0.0134).

1.3.3. Multinomials: a simple extension

A binomial probability involves 2 possible states (e.g., live or dead). What if there are more than 2 states? In this case, we use multinomial probabilities. As with our discussion of the binomial probability (above), we start by looking at the multinomial coefficient – the multinomial equivalent of the binomial coefficient. The multinomial is extremely useful in understanding the models we'll discuss in this book. The multinomial coefficient is nearly always introduced by way of a die tossing example. So, we'll stick with tradition and discuss this classic example here. You'll recall that a die has 6 sides - therefore 6 possible outcomes if your roll a die once. The multinomial coefficient corresponding to the 'die' example is

$$\binom{N}{n_1\ n_2\ n_3\ n_4\ n_5\ n_6} = \frac{N!}{n_1!n_2!n_3!n_4!n_5!n_6!} = \frac{N!}{\prod_{i=1}^{k} n_i!}$$

Note the use of the product operator '$\prod$' in the denominator. In a multinomial context, we assume that individual trials are independent, and that outcomes are mutually exclusive and all inclusive. Consider the 'classic' die example. Assume we throw the die 60 times ($N = 60$), and a record is kept of the number of times a 1, 2, 3, 4, 5 or 6 is observed. The outcomes of these 60 independent trials are shown below.

face	*number*	*notation*
1	13	y_1
2	10	y_2
3	8	y_3
4	10	y_4
5	12	y_5
6	7	y_6

Each trial has a mutually exclusive outcome (1 or 2 or 3 or 4 or 5 or 6). Note that there is a type of dependency in the cell counts in that once n and y_1, y_2, y_3, y_4 and y_5 are known, then y_6 can be obtained by subtraction, because the total (N) is known. Of course, the dependency applies to any count, not just y_6. This same dependency is also seen in the binomial case - if you know the total number of coin tosses, and the total number of heads observed, then you know the number of tails, by subtraction.

The multinomial distribution is useful in a large number of applications in ecology. Its probability function for $k = 6$ is

$$P(y_i \mid n, p_i) = \binom{n}{y_i} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5} p_6^{y_6}$$

Again, as was the case with the binomial probability, the multinomial coefficient does not involve any of the unknown parameters, and is conveniently ignored for many estimation issues.

This is a good thing, since in the simple die tossing example the multinomial coefficient is

$$\binom{n}{y_i} = \frac{60!}{13!10!8!10!12!7!}$$

which is an absurdly big number - beyond the capacity of your simple hand calculator to calculate. So, it is helpful that we can ignore it for all intents and purposes.

Some simple examples - suppose you role a 'fair' die 6 times (i.e., 6 trials), First, assume (y_1, y_2, y_3, y_4, y_5, y_6) is a multinomial random variable with parameters $p_1 = p_2 = \ldots p_6 = 0.1667$ and $N = 6$ What is the probability that each face is seen exactly once? This is written simply as:

$$
\begin{aligned}
P(1,1,1,1,1,1 \mid 6,1/6,1/6,1/6,1/6,1/6,1/6) &= \frac{6!}{1!1!1!1!1!1!}\left(\frac{1}{6}\right)^6 \\
&= \frac{5}{324} = 0.0154
\end{aligned}
$$

What is the probability that exactly four 1's occur, and two 2's occur in 6 tosses? In this case,

$$
\begin{aligned}
\mathcal{L}(4,2,0,0,0,0 \mid 6,1/6,1/6,1/6,1/6,1/6,1/6) &= \frac{6!}{4!2!0!0!0!0!}\left(\frac{1}{6}\right)^4\left(\frac{1}{6}\right)^2 \\
&= \frac{5}{15{,}552} \ll 0.0154
\end{aligned}
$$

As noted in our discussion of the binomial probability theorem, we are generally faced with the reverse problem – we do not know the parameters, but rather we want to estimate the parameters from the data. As we saw, these issues are the domain of the likelihood and log-likelihood functions. The key to this estimation issue is the multinomial distribution, and, particularly, the likelihood and log-likelihood functions

$$
\mathcal{L}(q \mid \text{data}) \quad \text{or} \quad \mathcal{L}(p_i \mid n_i, y_i)
$$

which we read as 'the likelihood of the parameters, given the data' – the left-hand expression is the more general one, where the symbol q indicates one or more parameters. The right-hand expression specifies the parameters of interest.

At first, the likelihood function looks pretty messy, but it is only a slightly different view of the probability function. Just as we saw from the binomial probability function, the multinomial function assumes N is given. The probability function further assumes that the parameters are given, while the likelihood function assumes the data are given. The likelihood function for the multinomial distribution is

$$
\mathcal{L}(p_i \mid n_i, y_i) = \binom{N}{y_i} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5} p_6^{y_6}
$$

Since the first term - the multinomial coefficient - is a constant, and since it doesn't involve any parameters, we ignore it. Next, because probabilities must sum to 1 (i.e., {sum of p_i over all i} = 1), there are only 5 'free' parameters, since the 6th one is defined by the other 5 (the 'dependency' issue we mentioned earlier), and the total, N. We will use the symbol K to denote the total number of estimable parameters in a model. Here, $K = 5$.

The likelihood function for $K = 5$, for example, is

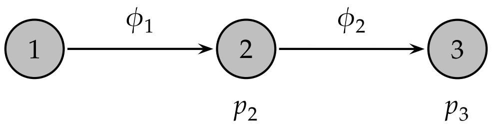
$$
\mathcal{L}(p_i \mid N, y_i) = \binom{N}{y_i} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5} \left(1 - \sum_{i=1}^{5} p_i\right)^{\left(N - \sum_{i=1}^{5} p_i\right)}
$$

So, just as we saw for the binomial example, we use a maximization routine (either analytical or numerical, depending on whether or not the likelihood can be expressed in closed form) to find the values of p_1, p_2, p_3, p_4 and p_5 that maximize the likelihood of the data that we observe. Remember - all we are doing is finding the values of the parameters which maximize the likelihood of observing the data that we see. Nothing more than that - at least conceptually.

1.4. Application to mark-recapture

Let's look at an example relevant to the task at hand (no more dice, or flipping coins.). Let's pretend we do a three year mark-recapture study, with 55 total marked individuals from a single *cohort*.* Once each year, we go out and look to see if we can 'see' (encounter) any of the 55 individuals we marked alive and in our sample. For now, we'll assume that we only encounter 'live' individuals.

The following represents the basic 'structure' of our sampling protocol.



In this diagram, each of the sampling events (referred to as 'sampling occasions') is indicated by a shaded grey circle. Our 'experiment' has three sampling occasions, numbered $1 \rightarrow 3$, respectively. In this diagram, time is moving forward going from left to right (i.e., sampling occasion 2 occurs one time step after sampling occasion 1, and so forth). Connecting the sampling occasions we have an arrow – the direction of the arrow indicates the direction of time - again, moving left to right, forward in time. We've also added two variables (symbols) to the diagram: ϕ and p. What do these represent?

For this example, these represent the two primary parameters which we believe (assume) govern the encounter process: ϕ_i (the probability of surviving from occasion i to $i+1$), and p_i (the probability that if alive and in the sample at time i, that the individual will be encountered). So, as shown on the diagram, ϕ_1 is the probability that an animal encountered and released alive at sampling occasion 1 will survive the interval from occasion $1 \rightarrow$ occasion 2, and so on. Similarly, p_2 is the probability that conditional on the individual being alive and in the sample, that it will be encountered at occasion 2, and so on. Why no p_1? Simple – p_1 is the probability of encountering a marked individual in the population, and none are marked prior to occasion 1 (which is when we start our study). In addition, the probability of encountering any individual (marked or otherwise) could only be calculated if we knew the size of the population, which we don't (this becomes an important consideration we will address in later chapters where we make use of estimated abundance). The important thing to remember here is the probability of being encountered at a particular sampling occasion is governed by two parameters: ϕ and p.

Now, as discussed earlier, if we encounter the animal, we record it in our data as '1'. If we don't encounter the animal, it's a '0'. So, based on a 3 year study, an animal with an encounter history of '111' was 'seen in the first year (the marking year), seen again in the second year, and also seen in the third year'. Compare this with an animal with an encounter history of '101'. This animal was 'seen in the first year, when it was marked, not seen in the second year, but seen again in the third year'. For a 3 occasion study, where the occasion refers to the sampling occasion, with a single release cohort, there are 4 possible encounter histories:

encounter history
111
101
110
100

* In statistics and demography, a *cohort* is a group of 'subjects' defined by experiencing a common event (typically birth) over a particular time span. In the present context, a cohort represents a group of individuals captured, marked, and released alive at the same point in time. These individuals would be part of the same *release cohort*.

Now, the key question we have to address, and (in simplest terms) the basis for analysis of data from marked individuals, is 'what is the probability of observing a particular encounter history?'. The probability of a particular encounter history is determined by a set of parameters - for this study, we know (or assume) that the parameters governing the probability of a given encounter history are ϕ and p. Based on the diagram on the previous page, we can write a probability expression corresponding to each of these possible encounter histories:

encounter history	*probability*
111	$\phi_1 p_2 \phi_2 p_3$
110	$\phi_1 p_2 (1 - \phi_2 p_3)$
101	$\phi_1 \,(1 - p_2)\, \phi_2 p_3$
100	$1 - \phi_1 p_2 - \phi_1 (1 - p_2) \phi_2 p_3$

For example, take encounter history '101'. The individual is marked and released on occasion 1 (the first 1 in the history), is not encountered on the second occasion, but is encountered on the third occasion. Now, because of this encounter on the third occasion, we know that the individual was in fact alive on the second occasion, but simply not encountered. So, we know the individual survived from occasion $1 \rightarrow 2$ (with probability ϕ_1), was not encountered at occasion 2 (with probability $1 - p_2$), and survived to occasion 3 (with probability ϕ_2) where it was encountered (with probability p_3). So, the probability of observing encounter history '101' would be $\phi_1(1 - p_2)\phi_2 p_3$.

Here are our 'data' - which consist of the observed frequencies of the 55 marked individuals with each of the 4 possible encounter histories:

encounter history	*frequency*
111	7
110	13
101	6
100	29

So, of the 55 individual marked and released alive in the release cohort, 7 were encountered on both sampling occasion 2 and sampling occasion 3, 13 were encountered on sampling occasion 2, but were not seen on sampling occasion 3, and so on.

The estimation problem, then, is to derive estimates of the parameters p_i and ϕ_i which maximizes the likelihood of observing the frequency of individuals with each of these 4 different encounter histories. Remember, the encounter histories are the data - we want to use the data to estimate the parameter values. What parameters? Again, recall also that the probability of a given encounter history is governed (in this case) by two parameters: ϕ, and p.

OK, so we've been playing with multinomials (above), and you might have suspected that these encounter data must be related to multinomial probabilities, and likelihoods. Good guess! The basic idea is to realize that the statistical likelihood of an actual encounter data set (as is tabulated above) is merely the product of the probabilities of the possible capture histories over those actually observed. As noted by Lebreton *et al*. (1992), because animals with the same encounter history have the same probability expression, then the number of individuals observed with each encounter history appears as an exponent of the corresponding probability in the likelihood. Thus, we write

$$\begin{aligned}\mathcal{L} =& (\phi_1 p_2 \phi_2 p_3)^{N_{(111)}} \times [\phi_1 p_2 (1-\phi_2 p_3)]^{N_{(110)}} \times [\phi_1 (1-p_2)\phi_2 p_3]^{N_{(101)}} \\ & \times [1-\phi_1 p_2 - \phi_1(1-p_2)\phi_2 p_3]^{N_{(100)}}\end{aligned}$$

where $N_{(ijk)}$ is the observed frequency of individuals with encounter history ijk.

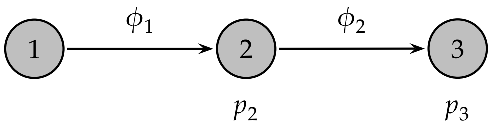
As with the binomial, we take the log transform of the likelihood expression, and after substituting the frequencies of each history, we get:

$$\begin{aligned}\ln \mathcal{L}(\phi_1, p_2, \phi_2, p_3) =& 7\ln(\phi_1 p_2 \phi_2 p_3) + 13\ln[\phi_1 p_2(1-\phi_2 p_3)] + 6\ln[\phi_1(1-p_2)\phi_2 p_3] \\ & + 29\ln[1-\phi_1 p_2 - \phi_1(1-p_2)\phi_2 p_3]\end{aligned}$$

All that remains is to derive the estimates of the parameters ϕ_i and p_i that maximize this likelihood.

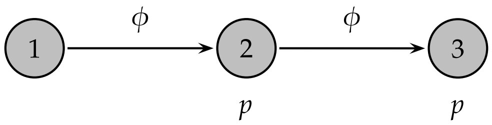
Let's go through a worked example, using the encounter history data tabulated on the preceding page. To this point, we have assumed that these encounter histories are governed by 'time-specific' variation in ϕ and p. In other words, we would write the probability statement for encounter history '111' as $\phi_1 p_2 \phi_2 p_3$.

These time-specific parameters are indicated in the following diagram:



Again, the subscripting indicates a different survival and recapture probability for each interval or sampling occasion.

However, what if instead we assume that the survival and recapture probabilities do not vary over time? In other words, $\phi_1 = \phi_2 = \phi$, and $p_2 = p_3 = p$. In this case, our diagram would now look like



What would the probability statements be for the respective encounter histories?

In fact, in this case deriving them is very straightforward – we simply drop the subscripts from the parameters in the probability expressions:

encounter history	*probability*
111	$\phi p \phi p$
110	$\phi p(1-\phi p)$
101	$\phi\,(1-p)\,\phi p$
100	$1-\phi p - \phi(1-p)\phi p$

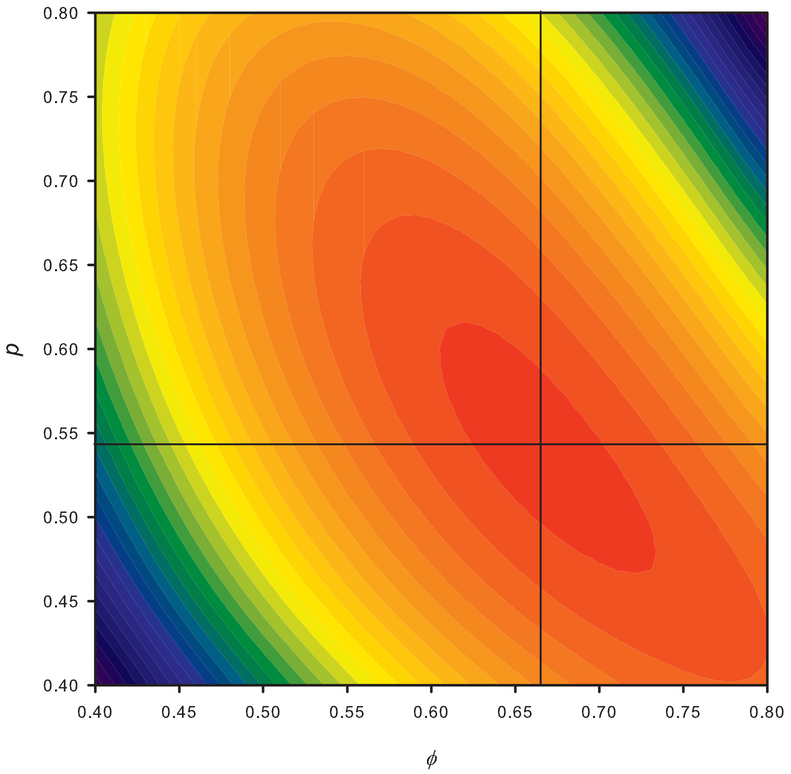
So, what would the likelihood look like? Well, given the frequencies, the likelihood would be:

$$\mathcal{L} = (\phi p \phi p)^{N^{111}} [\phi p(1-\phi p)]^{N^{110}} [\phi(1-p)\phi p]^{N^{101}} [1-\phi p - \phi(1-p)\phi]^{N^{100}}$$

Thus,

$$\ln \mathcal{L}(\phi, p) = 7 \ln(\phi p \phi p) + 13 \ln[\phi p(1-\phi p)] + 6 \ln[\phi(1-p)\phi p] + 29 \ln[1-\phi p - \phi(1-p)\phi p]$$

Again, we can use numerical methods to solve for the values of ϕ and p which maximize the likelihood of the observed frequencies of each encounter history. The likelihood profile for these data is plotted as a 2-dimensional contour plot, shown below:



We see that the maximum of the likelihood occurs at $p = 0.542$ and $\phi = 0.665$ (where the 2 dark black lines cross in the figure).

For this example, we used a numerical approach to find the MLE. In fact, for this example where ϕ and p are constant over time, the probability expressions are defined entirely by these two parameters, and we could (if we really had to) write the likelihood as two closed-form equations in ϕ and p, and derive estimates for ϕ and p analytically. All we need to do is (1) take the partial derivatives of the likelihood with respect to each of the parameters (ϕ, p) in turn $(\partial\mathcal{L}/\partial\phi, \partial\mathcal{L}/\partial p)$, (2) set each partial derivative to 0, and (3) solve the resulting set of simultaneous equations.

Solving simultaneous equations is something that most symbolic math software programs (e.g., **MAPLE**, **Mathematica**, **GAUSS**, **Maxima**) does extremely well. For this problem, the MLE are estimated analytically as $\hat{\phi} = 0.665$ and $\hat{p} = 0.542$ (just as we saw earlier using the numerical approach). However, recall that many of the likelihoods we'll be working with cannot be evaluated in closed form, so we will rely in numerical methods. Program **MARK** evaluates all likelihoods (and functions of likelihoods) numerically.

What is the actual value of the likelihood at this point? On the log scale, $\ln(\mathcal{L})$ is maximized at -65.041. For comparison, the maximized $\ln(\mathcal{L})$ for the model where both ϕ and p were allowed to vary with time is -65.035. Now, these likelihoods aren't very far apart - only in the second and third decimal places. Further, the two models (with constant ϕ and p, and with time varying ϕ and p) differ by only 1 estimable parameter (we'll talk a lot more about estimable parameters in coming lectures). So, a χ^2 test would have only 1 df. The difference in the $\ln(\mathcal{L})$ is 0.006 (actually, the test is based on $2\ln(\mathcal{L})$, so the difference is actually 0.012). This difference is not significant (in the familiar sense of 'statistical significance') at $P \gg 0.5$. So, the question we now face is, which model do we pick? This takes us to one of the main themes of this book – model selection. But, this is a glimpse of where we're headed.

1.5. Variance estimation for > 1 parameter

Earlier, we considered the derivation of the MLE and the variance for a simple situation involving only a single parameter. If in fact we have more than one parameter, the same idea we've just described for one parameter still works, but there is one important difference: a multi-parameter likelihood surface will have more than one second partial derivative. In fact, what we end up with a matrix of second partial derivatives, called the *Hessian*.

Consider for example, the log-likelihood of the simple mark-recapture data set we just analyzed in the preceding section:

$$\ln\mathcal{L}(\phi,p) = 7\ln(\phi p\phi p) + 13\ln[\phi p(1-\phi p)] + 6\ln[\phi(1-p)\phi p] + 29\ln[1-\phi p-\phi(1-p)\phi p]$$

Thus, the Hessian **H** (i.e., the matrix of second partial derivatives of the likelihood $\mathcal{L}$ with respect to ϕ and p) would be

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2\mathcal{L}}{\partial\phi^2} & \dfrac{\partial^2\mathcal{L}}{\partial\phi\partial p} \\ \dfrac{\partial^2\mathcal{L}}{\partial p\partial\phi} & \dfrac{\partial^2\mathcal{L}}{\partial p^2} \end{bmatrix}$$

We'll leave it as an exercise for you to derive the second partial derivatives corresponding to each of the elements of the Hessian. It isn't difficult, just somewhat cumbersome. For example,

$$\begin{aligned}\frac{\partial^2\mathcal{L}}{\partial\phi^2} = -\frac{26}{\phi^2} - \frac{26p}{\phi(1-\phi p)} - \frac{13[p(1-\phi p)-\phi p^2]}{\phi^2 p(1-\phi p)} + \\ \frac{13[p(1-\phi p)-\phi p^2]}{\phi(1-\phi p)^2} - \frac{58(1-p)p}{1-\phi p-\phi^2(1-p)p} - \frac{29[-p-2\phi(1-p)p]^2}{[1-\phi p-\phi^2(1-p)p]^2}\end{aligned}$$

Pretty ugly (and this for a simple model with only 2 parameters - ϕ and p - both of which are held constant over time in this example). Good thing **MARK** handles all this messy stuff for you.

Next, we evaluate the Hessian at the MLE for ϕ and p (i.e., we substitute the MLE values for our parameters - $\hat{\phi} = 0.6648$ and $\hat{p} = 0.5415$ - into the Hessian), which yields the information matrix $\mathbf{I}$

$$\mathbf{I} = \begin{bmatrix} -203.06775 & -136.83886 \\ -136.83886 & -147.43934 \end{bmatrix}$$

The negative inverse of the information matrix ($-\mathbf{I}^{-1}$) is the variance-covariance matrix for parameters ϕ and p

$$-\mathbf{I}^{-1} = -\begin{bmatrix} -203.06775 & -136.83886 \\ -136.83886 & -147.43934 \end{bmatrix}^{-1} = \begin{bmatrix} 0.0131 & -0.0122 \\ -0.0122 & 0.0181 \end{bmatrix}$$

Note that the variances are found along the diagonal of the matrix, while the off-diagonal elements are the covariances.

In general, for an arbitrary parameter θ, the <u>variance</u> of θ_i is given as the elements of the negative inverse of the information matrix corresponding to

$$\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_i}$$

while the <u>covariance</u> of θ_i with θ_j is given as the elements of the negative inverse of the information matrix corresponding to

$$\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j}$$

Obviously, the variance-covariance matrix is the basis for deriving measures of the precision of our estimates. But, as we'll see in later chapters, the variance-covariance matrix is used for much more - including estimating the number of estimable parameters in the model. While **MARK** handles all this for you, it's important to have a least a feel for what **MARK** is doing 'behind the scenes', and why. This will help with interpreting **MARK** output, and will help you figure out what might on occasion be causing problems. We will cover these and a few other related details in later chapters.

1.6. More than 'estimation' - ML and statistical testing

In the preceding, we focussed on the maximization of the likelihood as a means of deriving estimates of parameters and the sampling variance of those parameters. However, the other primary use of likelihood methods is for comparing the fits of different models.

We know that $\mathcal{L}(\hat{\theta})$ is the value of the likelihood function evaluated at the MLE $\hat{\theta}$, whereas $\mathcal{L}(\theta)$ is the likelihood for the true (but unknown) parameter θ. Since the MLE maximizes the likelihood for a given sample, then the value of the likelihood at the true parameter value θ is generally smaller than the MLE $\hat{\theta}$ (unless by chance $\hat{\theta}$ and θ happen to coincide).

This, combined with other properties of ML estimators noted earlier lead directly to several classic and general procedures for testing the statistical hypothesis that $H_o : \theta = \theta_0$. Here we briefly describe three of the more commonly used tests.

Fisher's Score Test

The 'score' is the slope of the log-likelihood at a particular value of θ. In other words, $S(\theta) = \partial \ln \mathcal{L}(\theta)/\partial\theta$. At the MLE, the score (slope) is 0 (by definition of a maximum).

Recall from earlier in this chapter that

$$\widehat{\text{var}}(\hat{\theta}) = \left[- \left(\frac{\partial^2 \ln \mathcal{L}(\theta \mid data)}{\partial \theta^2} \right) \right]^{-1}_{\theta=\hat{\theta}}$$

The term inside the inner parentheses is known as *Fisher information*

$$I(\theta) = -\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}$$

It can be shown that the score statistic

$$S_0 = \frac{S\,(\theta_0)}{\sqrt{I\,(\theta_0)}}$$

is asymptotically distributed as $N(0,1)$ under H_0.

Wald test

The Wald test relies on the asymptotic normality of the MLE $\hat{\theta}$. Given the normality of the MLE, we can calculate the test statistic

$$Z_0 = \frac{\hat{\theta} - \theta_0}{\sqrt{\widehat{\text{var}}(\hat{\theta})}}$$

which is asymptotically distributed as $N(0,1)$ under the null H_o.

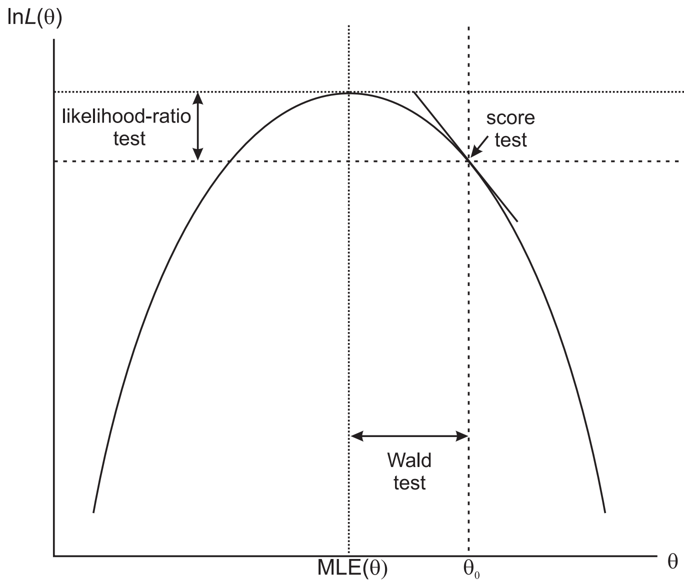
Likelihood ratio test

It is known that

$$2\left[\ln \mathcal{L}(\hat{\theta}) - \ln \mathcal{L}(\theta_0)\right]$$

follows an asymptotic χ^2 distribution with one degree of freedom.

The basic relationship among these tests is shown in the following diagram:



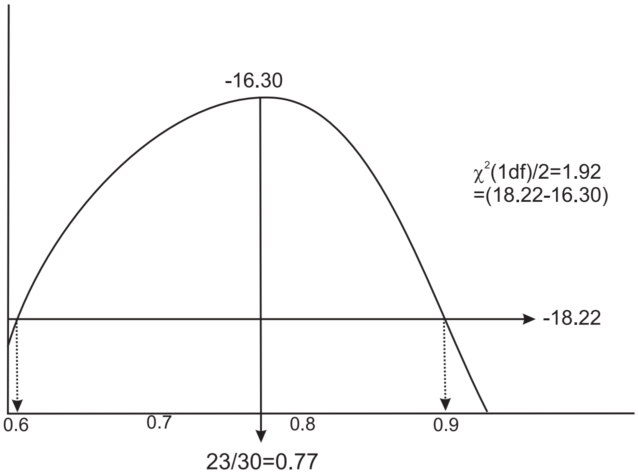
In general, these three tests are asymptotically equivalent, although in some applications, the score test has the practical advantage of not requiring the computation of the MLE at $\hat{\theta}$ (since S_0 depends only on the null value θ_0, which is specified in H_0). We consider one of these tests (the likelihood ratio test) in much more detail in Chapter 4.

1.7. Technical aside: a bit more on variances

As we discussed earlier, the classic MLE approach to variance calculation (for purposes of creating a SE and so forth) is to use the negative inverse of the 2nd derivative of the MLE evaluated at the MLE. However, the problem with this approach is that, in general, it leads to derivation of symmetrical 95% CI, and in many cases – especially for parameters that are bounded [0, 1] – this makes no sense. A simple example will show what we mean. Suppose we release 30 animals, and find 1 survivor. We know from last time that the MLE for the survival probability is $(1/30) = 0.03333$. We also know from earlier in this chapter that the classical estimator for the variance, based on the 2nd derivative, is

$$\widehat{\text{var}}(\hat{p}) = \frac{p(1-p)}{N} = \frac{0.0333(1-0.0333)}{30} = 0.0010741$$

So, based on this, the 95% CI using classical approaches would be ±1.96(SE), where the SE (standard error) is estimated as the square-root of the variance. Thus, given $\widehat{\text{var}} = 0.001074$, the 95% CI would be ±1.96(0.03277), or [0.098, −0.031]. OK, so what's wrong with this? Well, clearly, we don't expect a 95% CI to ever allow values < 0 (or > 1) for a parameter that is logically bounded to fall between 0 and 1 (like ϕ or p). So, there must be a problem, right? Well, somewhat. Fortunately, however, there is a better way, using something called the *profile likelihood* approach, which makes more explicit use of the shape of the likelihood. We'll go into the profile likelihood in further detail in later chapters, but to briefly introduce the concepts - consider the following diagram, which shows the maximum part of the log likelihood for ϕ, given $N = 30$, $y = 23$ (i.e., 23/30 survive).



Profile likelihood confidence intervals are based on the log-likelihood function. For a single parameter, likelihood theory shows that the 2 points 1.92 units down from the maximum of the log likelihood function provide a 95% confidence interval when there is no extra-binomial variation (i.e., $c = 1$; see Chapter 5). The value 1.92 is half of the $\chi^2_1 = 3.84$. Thus, the same confidence interval can be computed with the *deviance* by adding 3.84 to the minimum of the deviance function, where the deviance is the log-likelihood multiplied by -2 minus the -2 log likelihood value of the saturated model (more and these concepts in later chapters).

Put another way, we use the critical value of 1.92 to derive the *profile* - you take the value of the log likelihood at the maximum (for this example, the maximum occurs at -16.30), add 1.92 to it (yielding -18.22; note we keep the negative sign here), and look to see where the -18.22 line intersects with the *profile* of the log likelihood function. In this case, we see that the intersection occurs at approximately 0.6 and 0.9. The MLE is $(23/30) = 0.767$, so clearly, the profile 95% CI is not symmetrical around this MLE value. But, it **is** bounded on the interval $[0, 1]$. The profile likelihood is the preferred approach to deriving 95% CI. The biggest limit to using it is computational – it simply takes more work to derive a profile likelihood (and corresponding CI). Fortunately, **MARK** does all the work for us.

1.8. Summary

That's it for Chapter 1! Nothing about **MARK**, but some important background. Beginning with Chapter 2, we'll consider formatting of our data (the 'encounter histories' we introduced briefly in this chapter). After that, the real details of using program **MARK**. Our suggestion at this stage is to (i) leave your own data alone - you need to master the basics first. This means working through at least chapters 3 $\rightarrow$ 8, in sequence, using the example data sets. Chapter 9 and higher refer to specific data types - one of more may be of particular interest to you. Then, when you're ready (i.e., have a good understanding of the basic concepts), (ii) get <u>your</u> data in shape - this is covered in Chapter 2.