Lecture 8 -- Open Population Capture-Recapture Models

This is a rich class of models that are both quite similar to the band/tag recovery models and also have some important differences. Key references include Pollock et al. (1990) for the Jolly-Seber model and Lebreton et al. (1992) for the Cormack-Jolly-Seber model. We will assume the reader has some familiarity with these source materials. Many will find Jolly (1965) very informative. This paper was a landmark and contained several insights that were not fully appreciated until 20 years latter. Seber (1982) also has a nice discussion of the basic, time-specific model.

Animals are captured on *k* occasions (say, years) and given a unique mark during a relatively short tagging period (say, week) each year. Time periods could also be weeks, months, or multiple year intervals. After occassion 1, both marked and unmarked animals are caught; tag numbers of the marked animals are recorded and the unmarked animals are marked. Animals are released back into the population; accidental deaths (losses on capture) are allowed. The Jolly-Seber model (after Jolly 1965 and Seber 1965) is fairly general and serves as a starting point for open C-R modeling (see White et al (1982, chapter 8). This model allows year-specific estimates of apparent survival (ϕ), capture probability (*p*), population size (*N*), and the number of new individuals entering the population (*B*). While population size can be estimated, it is often very difficult to avoid substantial bias in the estimation of this parameter set because of individual heterogeneity and other issues. Likewise, the B_j are subject to bias and often terrible precision. Program *POPAN7* is the most capable software for the analysis of J-S data where there is interest in estimation of population size and “births."

The Cormack-Jolly-Seber model is a restricted model (so named after Cormack's model appeared in 1964) and allows only year-specific estimates of ϕ and *p*. While less general, it has proven to be the more useful model for several reasons. Program *MARK* handles C-J-S data very well and that will be the emphasis here. The material in Lebreton et al. (1992) is current (except perhaps pages 80-84) and thus the notes provided here represent only a brief overview. Students are strongly encouraged to study Lebreton et al. and the examples provided. One important biological issue is that only apparent survival can be estimated in the open C-R studies; that is 1–ϕ represents both animals that died and animals that merely left the population (emigration). In general, $\phi \leq S$. This is often a significant matter and often misunderstood.

Unlike the band recovery models where “sampling" tends to take place throughout the range of the marked population, C-R sampling tends to be done by the investigators. Thus, animals are marked in a relatively small locality each year for several years. The only way recaptures are made is for these animals to come back to the same (relatively very small) area where capturing is being conducted. Thus, an animal that comes back to the same general area may not be recapture as it is a mile or two away from the capture site. This issue can be very problematic with migratory birds or fish that move considerable distances during the course of a year. Still, there are often cases where there is biological interest in ϕ and the fact that some animals merely “left" is not problematic to the interpretation of the data. White et al. (1982: 183) provide some useful figures on the “open" models.

Assumptions of the Open C-R Models

1. Tagged fish are representative of the population of fish to which inferences are to be made (i.e., tagged fish are a random sample of the population of interest).

2. Numbers of releases are known (the R_i).

3. Tagging is accurate, no tag loss, no misread tags, no data entry errors.

4. All releases are made within relatively brief time periods. (Relative to time intervals between tagging periods.)

5. The fate of individual fish and the fates of fish in differing cohorts are independent. (counter example: banding of a mated pair of adult Canada geese; these might behave as one unit, instead of two independent trials (flips)).

6. Fish in an identifiable class or group have the same survival and reporting probability (parameter homogeneity) for each time interval.

7. Parameter estimates and sampling covariances are based on a good approximating model. The basis for inference is the model, thus, results are conditional on the model used.

Known Constants:

R_i The number of animals released in year *i*. These released are typically made up of animals newly captured, tagged and released, as well as animals already tagged and re-released into a new cohort.

Random Variables:

m_{ij} The number of recaptures in year *j* from releases in year *i*. A matrix of first recaptures.

Parameters:

ϕ_j Conditional probability of apparent survival in year *j*, given the animal is alive at the beginning of year *j*. ϕ_j relates to the interval between capture periods!

p_j Conditional probability of being captured (or recaptured) in year *j*, given the animal was alive at the beginning of year *j*.

We use *K* to denote the total number of estimable parameters in a model.

It is assumed that capture and recapture is done within a relatively short time period (say, 1-3 weeks if survival is to be estimated over an annual period.

In general, the occasions can be separated by days or weeks in the case of short-lived insects or months or small mammals; or years or groups of years. Here, we will often use the word "year."

The parameter p_j, the capture probability at occasion *j*, can be an "encounter" including physical capture or the resighting of marked individuals. Let the number of animals released at each occasion be R_i and the recapture data be summarized as an *m* array, m_{ij}. The *m* array provides the biologist with a way to view the data and a compact form. The data in this array are conditional on last release. Thus, only the first (re)capture following release is summarized in the *m* array. The *m* array is useful in estimation, but a finer summary of the data is needed for a full goodness-of-fit test (i.e., either the encounter history matrix or the full *m* array).

Open population C-R data are summarized by a encounter history (EH) matrix. A full discussion of this is found in Burnham et al. (1987:28-29). Here a **1** denotes capture (or recapture) and a **0** denotes not captured (or recaptured) at occasion *j*. For example, consider a C-R survey over 7 occasions. A particular animal might have the following encounter history,

{1001101}

meaning that it was initially captured (and given a uniquely numbered tag) on occasion 1, was not caught on occasion 2, was recaptured and re-released on occasions 3, 4, and 5, not captured on occasion 6, and recaptured on occasion 7. To better see the relationship between the encounter history and the parameters under model $\{\phi_t\, p_t\}$, examine the following diagram.

$$\begin{array}{ccccccccccccc} 1 & & 0 & & 0 & & 1 & & 1 & & 0 & & 1 \\ & \phi_1 & & \phi_2 & & \phi_3 & & \phi_4 & & \phi_5 & & \phi_6 & \\ & & p_2 & & p_3 & & p_4 & & p_5 & & p_6 & & p_7 \end{array}$$

Note that no capture probability is associated with the first release, because the animal is inserted into the population at this time. The probability of observing this encounter history is $\phi_1(1-p_2)\phi_2(1-p_3)\phi_3 p_4 \phi_4 p_5 \phi_5 (1-p_6)\phi_6 p_7$.

The following example shows an animal that is first uniquely marked on the 3rd occasion.

$$\begin{array}{ccccccccccccc} 0 & & 0 & & 1 & & 1 & & 1 & & 0 & & 1 \\ & & & & & \phi_3 & & \phi_4 & & \phi_5 & & \phi_6 & \\ & & & & & & p_4 & & p_5 & & p_6 & & p_7 \end{array}$$

No information on ϕ_1 and ϕ_2, or p_2 and p_3 is provided by this animal. The probability of observing this encounter history is $\phi_3 p_4 \phi_4 p_5 \phi_5 (1-p_6)\phi_6 p_7$.

A more complicated probability results when the animal is not captured on the last occasion(s). Consider the encounter history 0011110 that results in the following probability: $\phi_3 p_4 \phi_4 p_5 \phi_5 p_6 [\phi_6(1 - p_7) + (1 - \phi_6)]$. The last term in brackets is the probability that the animal is alive and not captured, plus the probability that the animal died before occasion 7. If an animal is not observed for the last 2 occasions, giving an encounter history of 0011100, the following probability results: $\phi_3 p_4 \phi_4 p_5 \{\phi_5(1 - p_6)[\phi_6(1 - p_7) + (1 - \phi_6)] + (1 - \phi_5)\}$. As you can see, these encounter history probabilities get quite complicated when animals are not observed for several occasions at the end of a history. Note that if the animal is not released back into the population, the complexity disappears, because the animals is known to not be available for encounters.

The input data file to Program MARK or Program RELEASE could be structured so that every animal is represented by a row in the encounter history matrix. This is often done.

A convenient alternative is to count all the animals with the same encounter history (say, {00101} and note that 17 animals had this exact encounter history. Then, the input file to Program MARK can show this as

```
00101  17;
```

to indicate to *MARK* that 17 animals share this EH, rather than having to enter 17 lines into the input file. This method is handy when there are 2 or more groups (say, males and females or treatment and control). Then, one might enter

```
00101  17 22;
```

to denote that 17 males and 22 females had the exact EH {00101}. The Cormack-Jolly-Seber model allows for “losses on capture" whereby animals might be deliberately removed from the study population or might accidentally die in the trap. Thus, the number captured may differ from the number re-released. The number lost on capture is denoted with a minus sign, e.g.,

```
00101 -1 -3;
```

to indicate that 1 male and 3 females were not re-released after capture at occasion 5.

There is the concept of a EH matrix for band recovery data, but in that case the histories are of a few simple types;

$\{100 \cdots 01\}$ or $\{100 \cdots\}$,

that is, marked animals are either reported (dead) once following initial banding and release, or they are never reported following initial release. This simplicity makes it compelling to merely summarize the data in the m_{ij} array. The matter is more complicated with open population C-R data. Here, one can consider the EH matrix as the basic data and estimation and some testing can be done with data in this format. In addition, there are two ways to summarize the C-R data: the full m_{ij} array and the reduced m_{ij} array. A full discussion of

these concepts is given in Burnham et al. 1987:34-35). The full *m* array allows efficient goodness-of-fit testing and estimation. We will cover this issue at a latter time but mention that program *RELEASE* is still quite useful is providing a summary in the form of a full or reduced *m* array, and providing detailed goodness-of-fit testing (its estimation capabilities are more limited).

A major difference between band recovery and open C-R models is in the expectations of the data. Consider first the expectation, under model $\{\phi, p\}$ for the encounter history {1001101}. Although there are 7 values entered in the encounter history,

$$\phi(1-p)\phi(1-p)\phi p\phi p\phi(1-p)\phi p.$$

The pattern can be more easily seen if we denote $(1-p)$ as q and then use parentheses to mark pairs of ϕ and q,

$$(\phi q)(\phi q)(\phi p)(\phi p)(\phi q)(\phi p).$$

Under model $\{\phi_t, p\}$ the expectation is

$$(\phi_1 q)(\phi_2 q)(\phi_3 p)(\phi_4 p)(\phi_5 q)(\phi_6 p).$$

In C-R data it is essential to formally include the fact that animals alive at the beginning of occasion *i* were captured (p) or not ($1-p$).

The full and reduced *m* arrays can be constructed (*MARK* does this) from the EH matrix but one cannot construct the EH matrix from the reduced *m* array. Modeling can be done in some cases from either an *m* array or the EH matrix, but modeling individual covariates must be based on the EH matrix.

The m_{ij} array also can be expressed as expectations under various models. For example,

	2	3	4	5
R_1	m_{12}	m_{13}	m_{14}	m_{15}
R_2		m_{23}	$\boldsymbol{m_{24}}$	m_{25}
R_3			m_{34}	m_{35}
R_4				m_{45} .

The element m_{24} denotes the number of marked individuals initially released at time 2 that were *first* recaptured (encountered) at time 4. In fact, some of those released at time 2 might have been initially banded at time 1 (part of R_1), caught at time 2 (and thus appearing in m_{12}) and re-released at time 2 (part of R_2). Both of the encounter histories {1101...} and {0101...} would contribute to the element m_{24}.

Thus, it is clear that a single individual appears in a row of the m_{ij} array only once (thus, the multinomial might serve as a useful approximating model). Of course, a final column could have been shown to indicate the number of animals never recaptured from a released cohort. Expectations of the m_{ij} array also have a major difference from the expectations of band recovery data. For example, under model $\{\phi_t, p_t\}$

C-R data $\quad E(m_{24}) = R_2\phi_2(1-p_3)\phi_3 p_4$

Band recovery data $\quad E(m_{24}) = R_2 S_2 S_3(1-S_4)r_4$.

The primary difference is that the **probability of not being caught at time *j* (i.e., 1–*p*)** must also be included in the expectations. Another example will make this more clear;

$$E(m_{16}) = R_1\phi_1(1-p_2)\phi_2(1-p_3)\phi_3(1-p_4)\phi_4(1-p_5)\phi_5 p_6 .$$

This denotes that the animal survived years 1, 2, 3, 4, 5, and was first recaptured in year 6. However, it is also necessary to clearly indicate that the individual was not caught in years 2, 3, 4, and 5.

Other differences include the indexing and a switch from survival (S) to apparent survival (ϕ). Program *MARK* makes this somewhat transparent in modeling the data. Thus, the PIMs can be set up in a manner similar to that for recovery data. Likewise, modeling survival or capture probabilities as functions of covariate via link functions is similar. Model selection theory and other tools are the same as those for the open C-R models.

Generally, the input data from open C-R studies is a encounter history matrix (see Lebreton et al. 1992, Table 2). *MARK* can build the *m* array from this matrix. Lebreton et al. (1992:71) gives other examples of expectations, mostly involving the encounter history matrix for individuals.

Why Model?

The model links, in a formal manner, the data $\{R_i, m_{ij}\}$, assumptions, unknown parameters (the ϕ_j and p_j) and allows rigor in making inductive inferences via likelihood and information theory.

Models are an essential component of science.

Problems occur if an estimate $\hat{\phi}_j > 1$; this is a diagnostic indicating an over-parameterized model. The model assuming the apparent survival and recapture probabilities $\{\phi, p\}$ are constant and always have estimates in the $\{0, 1\}$ range.

Note, that estimated ϕ is an apparent survival probability in most cases because 1–ϕ includes both animals that left the population through emigration as well as those that died. ϕ is a finite rate (not instantaneous).

Models for Open Population Capture-Recapture Data

Modeling open population capture-recapture data involves *reparameterizing* the multinomial likelihood and log-likelihood. First, we look at the structure of the parameters that might be hypothesized to underly recovery data that we observe.

Consider the following m_{ij} array for the sphinx moth –

i	R_i	m_{ij}			
		j=2	3	4	5
1	800	30	60	117	**44**
2	730		55	91	51
3	715			119	56
4	807				70
5	601				

We assume the occasions are a day and the interval between occasions are weeks in this example. More often, the occasions are in annual increments. Notice that 44 moths tagged at the beginning of week 1 were *first* recaptured in week 5; m_{15}, shown in bold. Their encounter history would be {10001}.

The number released at time 2 (R_2) was 730 moths; this is 700 newly captured moths plus 30 that were first recaptured at occasion 2 and re-released into cohort 2 (the R_2). In a similar manner, 715 moths were released at occasion 3; 600 of these were newly captured moths and 115 were moths re-released from being released in cohorts 1 (60 moths) and 2 (55 moths).

Each row of the array (recapture matrix) is a multinomial distribution (under certain assumptions). Moths released in the first cohort and *first* recaptured in week (or occasion) 2, ..., 5 or “never." The subscripting differs from that used in modeling band recovery data. Note, also, that once a moth is first recaptured, it is either “lost on capture" or re-released into another cohort (it becomes part of another R_i).

We begin with the saturated model for the moth data. The **Saturated Model** has as many parameters as there are cells. Here the subscripts *i* and *j* denote week-specific parameters (the *j*), specific to each released cohort (the *i*).

Number Released	**Expected Number of First Recaptures Following Release**			
	j=2	3	4	5
R_1	$R_1\theta_{12}$	$R_1\theta_{13}$	$R_1\theta_{14}$	$\boldsymbol{R_1\theta_{15}}$
R_2		$R_2\theta_{23}$	$R_2\theta_{24}$	$R_2\theta_{25}$
R_3			$R_3\theta_{34}$	$R_3\theta_{35}$
$\boldsymbol{R_4}$				$\boldsymbol{R_4\theta_{45}}$
R_5				

Thus, K = 10 as there are 10 cells in the moth example. The fit of the saturated model is perfect and serves as a basis for comparison of fit of other models; it is also needed to define deviance of other models (i.e., models with fewer parameters).

Model $\{\phi_t, p_t\}$

Here the subscript *t* denotes week-specific survival and recapture probabilities.

Number Released	**Expected Number of First Recaptures following Release**			
	j=2	3	4	5
R_1	$R_1\phi_1p_2$	$R_1\phi_1(1-p_2)\phi_2p_3$	$R_1\phi_1(1-p_2)\phi_2(1-p_3)\phi_3p_4$	$R_1\phi_1(1-p_2)\phi_2(1-p_3)\phi_3(1-p_4)\phi_4p_5$
R_2		$R_2\phi_2p_3$	$R_2\phi_2(1-p_3)\phi_3p_4$	$R_2\phi_2(1-p_3)\phi_3(1-p_4)\phi_4p_5$
R_3			$R_3\phi_3p_4$	$R_3\phi_3(1-p_4)\phi_4p_5$
R_4				$R_4\phi_4p_5$
$\boldsymbol{R_5}$				

Model $\{\phi_t, p\}$ Note the p appears without a subscript, meaning this parameter is constant and survival probability is time-specific.

Number Released	**Expected Number of First Recaptures following Release**			
	j=2	3	4	5
R_1	$R_1\phi_1 p$	$R_1\phi_1(1\text{-}p)\phi_2 p$	$R_1\phi_1(1-p)\phi_2(1\text{-}p)\phi_3 p$	$R_1\phi_1(1-p)\phi_2(1-p)\phi_3(1\text{-}p)\phi_4 p$
R_2		$R_2\phi_2 p$	$R_2\phi_2(1\text{-}p)\phi_3 p$	$R_2\phi_2(1-p)\phi_3(1-p)\phi_4 p$
R_3			$R_3\phi_3 p$	$R_3\phi_3(1\text{-}p)\phi_4 p$
R_4				$R_4\phi_4 p$
$\boldsymbol{R_5}$				

Model $\{\phi, p_t\}$ In this case, apparent survival is constant, but recapture probability is week-specific.

Number Released	**Expected Number of First Recaptures following Release**			
	j=2	3	4	5
R_1	$R_1\phi p_2$	$R_1\phi(1\text{-}p_2)\phi p_3$	$R_1\phi(1-p_2)\phi(1\text{-}p_3)\phi p_4$	$R_1\phi(1-p_2)\phi(1-p_3)\phi(1\text{-}p_4)\phi p_5$
R_2		$R_2\phi p_3$	$R_2\phi(1\text{-}p_3)\phi p_4$	$R_2\phi(1-p_3)\phi(1-p_4)\phi p_5$
R_3			$R_3\phi p_4$	$R_3\phi(1\text{-}p_4)\phi p_5$
R_4				$R_4\phi p_5$
$\boldsymbol{R_5}$				

Model $\{\phi, p\}$ Note, here neither parameter is subscripted, implying they are constants (no time effects).

Number Released	**Expected Number of First Recaptures following Release**			
	j=2	3	4	5
R_1	$R_1\phi p$	$R_1\phi(1-p)\phi p$	$R_1\phi(1-p)\phi(1-p)\phi p$	$R_1\phi(1-p)\phi(1-p)\phi(1-p)\phi p$
R_2		$R_2\phi p$	$R_2\phi(1-p)\phi p$	$R_2\phi(1-p)\phi(1-p)\phi p$
R_3			$R_3\phi p$	$R_3\phi(1-p)\phi p$
R_4				$R_4\phi p$
$\boldsymbol{R_5}$				

The moth data could have been shown with the number <u>never</u> <u>recaptured:</u>

i	R_i	m_{ij}				
		j=2	**3**	**4**	**5**	**never**
1	**800**	**30**	**60**	**117**	**44**	549
2	730		55	91	51	**533**
3	715			119	56	540
4	807				70	737
5	601					

The number never recaptured from the cohort released in week 2 is 533 and its expectation under Model $\{\phi_t, p_t\}$ is:

$$R_2 - \left(R_2\phi_2 p_3 + R_2\phi_2(1-p_3)\phi_3 p_4 + R_2\phi_2(1-p_3)\phi_3(1-p_4)\phi_4 p_5\right)$$

or

$$730 - (55 + 91 + 51) = 533.$$

The expected number never recaptured from the cohort released in week 2 varies by model;

e.g., for Model $\{\phi, p\}$ it is

$$R_2 - \left(R_2\phi p + R_2\phi(1\text{-}p)\phi p + R_2\phi(1\text{–}p)\phi(1\text{–}p)\phi p\right)$$

and this could be computed given the parameters ϕ and p, but in general, it will not equal 533 except for model $\{\phi_t, p_t\}$. Thus, there is goodness-of-fit information in the “never recaptured" cell for models other than $\{\phi_t, p_t\}$.

Program *MARK* computes these values, however, one must be aware of this additional cell in the multinomial distribution.

For modeling in a likelihood framework, we will need to convert expectations to *probabilities.* These can be obtained by merely dividing the expected cell values by the number released.

Model $\{\phi_t, p_t\}$

Number Released	**Matrix of Cell Probabilities for m_{ij} E(m_{ij}/R_i)**			
R_1	$\phi_1 p_2$	$\phi_1(1-p_2)\phi_2 p_3$	$\phi_1(1-p_2)\phi_2(1-p_3)\phi_3 p_4$	$\phi_1(1-p_2)\phi_2(1-p_3)\phi_3(1-p_4)\phi_4 p_5$
R_2		$\phi_2 p_3$	$\phi_2(1-p_3)\phi_3 p_4$	$\phi_2(1-p_3)\phi_3(1-p_4)\phi_4 p_5$
R_3			$\phi_3 p_4$	$\phi_3(1-p_4)\phi_4 p_5$
R_4				$\phi_4 p_5$
$\boldsymbol{R_5}$				

Model $\{\phi, p\}$

Number Tagged	**Matrix of Cell Probabilities for m_{ij} E(m_{ij}/R_i)**			
R_1	ϕp	$\phi(1-p)\phi p$	$\phi(1-p)\phi(1-p)\phi p$	$\phi(1-p)\phi(1-p)\phi(1-p)\phi p$
R_2		ϕp	$\phi(1-p)\phi p$	$\phi(1-p)\phi(1-p)\phi p$
R_3			ϕp	$\phi(1-p)\phi p$
R_4				ϕp
$\boldsymbol{R_5}$				

We will switch back and forth between expected values and probabilities, so people need to get familiar with both expressions.

Consider the $\log_e \mathcal{L}(\phi, p \mid R_1, m_{1j})$ for the first week of release of tagged sphinx moths. This is the log of the likelihood of the parameters (the constant survival probabilities ϕ and p), given the data from the release of tagged moths in week 1 (the R_1 and m_{1j})

Consider the original data set on sphinx moths and the first row of the m_{ij} matrix;

i	R_i	m_{ij}			
		j=2	3	4	5
1	**800**	**30**	**60**	**117**	**44**
2	730		55	91	51
3	715			119	56
4	807				70
5	601				

The cell probabilities under Model **$\{\phi, p\}$** are:

Number Tagged	**Matrix of Cell Probabilities for m_{ij}**			**E(m_{ij}/R_i)**
R_1	ϕp	$\phi(1-p)\phi p$	$\phi(1-p)\phi(1-p)\phi p$	$\phi(1-p)\phi(1-p)\phi(1-p)\phi p$
R_2		ϕp	$\phi(1-p)\phi p$	$\phi(1-p)\phi(1-p)\phi p$
R_3			ϕp	$\phi(1-p)\phi p$
R_4				ϕp
R_5				

The the log-likelihood of the parameters ϕ and p, *given* the data (the number released in week 1 and the number of first recaptures in week j **from those released in week 1).**

$$\log_e \mathcal{L}_1(\phi, p \mid R_1, m_{1j}) = m_{12}\log_e\left(\phi p\right) + m_{13}\log_e\left(\phi(1-p)\phi p\right) + m_{14}\log_e\left(\phi(1-p)\phi(1-p)\phi p\right) +$$

$$m_{15}\log_e\left(\phi(1-p)\phi(1-p)\phi(1-p)\phi p\right)$$

+ (a term for those never recaptured, *MARK* handles this)

The log-likelihood function for the tag recoveries from releases in week 2 is similar,

$$\log_e \mathcal{L}_2(\phi, p \mid R_2, m_{2j}) = m_{23}\log_e\left(\phi p\right) + m_{24}\log_e\left(\phi(1-p)\phi p\right) +$$

$$m_{25}\log_e\left(\phi(1-p)\phi(1-p)\phi p\right) + \text{“never"}$$

The log-likelihood functions for the tag recoveries from releases in weeks 3 and 4 are

$$\log_e \mathcal{L}_3(\phi, p \mid R_3, m_{3j}) = + m_{34}\log\left(\phi p\right) + m_{35}\log\left(\phi(1-p)\phi p\right)$$

$$\log_e \mathcal{L}_4(\phi, p \mid R_4, m_{4j}) = m_{45}\log\left(\phi p\right) .$$

Example, the log-likelihood for releases in week 3:

$$\log_e \mathcal{L}_3(\phi, p \mid 715, 119, 56) = 119\log\left(\phi p\right) + 56\log\left(\phi(1-p)\phi p\right)$$

$$+ \mathbf{(715-119-56)\log}\left(\mathbf{1-}\left(\boldsymbol{\phi p}\right) + \left(\boldsymbol{\phi(1-p)\phi p}\right)\right)$$

This function assumes the “data" are given; indeed, the data here include the 715 moths released at the beginning of week 3 and the number first recaptured in weeks 3 and 4. Only the 2 parameters (ϕ and p) are unknown and the objects of interest. The log-likelihood function here is a function only of ϕ and p (note, $K = 2$).

Note the final term for those moths never recaptured from the 715 released is shown explicitly in the final expression (above, in bold).

The total log-likelihood for all the recapture data for the 5 weeks of release is merely the sum of the individual log-likelihoods (assuming independence of released cohorts),

$$\mathbf{\log_e \mathcal{L}(\phi, p \mid R_i, m_{ij}) = \log_e(\mathcal{L}_1) + \log_e(\mathcal{L}_2) + \ldots + \log_e(\mathcal{L}_5).}$$

Thus,

$$\log_e \mathcal{L}(\phi, p \mid R_i, m_{ij}) = m_{12}\log_e\left(\phi p\right) + m_{13}\log_e\left(\phi(1\text{-}p)\phi p\right) + m_{14}\log_e\left(\phi(1-p)\phi(1\text{-}p)\phi p\right) +$$

$$m_{15}\log_e\left(\phi(1-p)\phi(1-p)\phi(1\text{-}p)\phi p\right)$$

$$+ \quad m_{23}\log_e\left(\phi p\right) + m_{24}\log_e\left(\phi(1-p)\phi p\right) + m_{25}\log_e\left(\phi(1-p)\phi(1-p)\phi p\right)$$

$$+ \quad m_{34}\log_e\left(\phi p\right) + m_{35}\log_e\left(\phi(1\text{-}p)\phi p\right)$$

$$+ \quad m_{45}\log_e\left(\phi p\right) .$$

+ complex “never recovered" terms.

Note, each term involving parameters in the log-likelihood function is of the form

DATA * LOG(PROBABILITY).

The “**DATA**" are the R_i (number tagged in week *i*) and the m_{ij} (number of first recaptures following release at occasion *i*.

The “**PROBABILITY**" of observing the data is some expression of the unknown, underlying *parameters*, given a particular *model* (the ϕ_j and the p_j). These are often called “cell probabilities."

The log likelihood function can also be constructed from the encounter histories matrix, which is how *MARK* actually works. Thus, the log of the probability of each encounter history is multiplied by the number of animals with that history, i.e.,

$$\log_e \mathcal{L}(\phi, p \mid \text{encounter histories}) =$$

$$\sum_{\text{over all encounter histories}} (\text{Number of animals}) * \log_e\left(\text{Probability(Encounter History)}\right) .$$

As you would expect, the resulting likelihood is identical for the likelihood constructed from the m_{ij} array, although proving this result to yourself will require considerable algebra. An important implication of this result is that the underlying multinomial distribution from which

the likelihood function is developed can be constructed based on encounter histories, i.e., each animal is placed in a multinomial cell determined from its encounter history.

Little complications that are important (trust *MARK*):

1. A multinomial cell must be included for individuals in each released cohort that are "never" recaptured.

2. The multinomial coefficient $\binom{n}{y_i}$ is shorthand for

$$n!/\left((y_1)!\,(y_2)!\,\cdot\,\cdot\,\cdot\,(y_k)!\right),$$

This term does not involve any of the unknown parameters and is ignored for many estimation issues. *MARK* handles this issue.

In the moth tagging data we have multinomial coefficients in the overall likelihood; each of the form

$$\binom{R_i}{m_{ij}}\quad \text{, where } i = 1, ..., 4 \text{ and } j = 2, ..., 5.$$

For $i = 1$,

$$R_1!\Big/\left((m_{12})!\,(m_{13})!\,\cdot\,\cdot\,\cdot\,(m_{15})!\,(R_1-\textstyle\sum m_{1j})!\right)$$

or

$$800!\Big/\left(30!\;60!\;117!\;44!\;(800–251)!\right).$$

You can see why it is convenient to ignore this term! Program *DERIVE* does such things easily; we will get to it soon.

Some Questions:

1. What if some recaptured individuals are deliberately removed on occasion *j*?
What is the likely effect of the estimators ϕ_i?

2. What if the sample size R_i was actually made up of *m* groups of brood mates. What might one worry about (in terms of the multinomial model)?

3. The total log-likelihood is a sum of the log-likelihoods for each of the released cohorts. But this is based on the notion that the cohorts are independent. Is this a good assumption? Why? Why not? When might it fail?

Why is the literature on the Jolly-Seber model so hard to follow? Why so much notation that seems not to be covered in lectures? Why is the log-likelihood function not more prominent?

The original work on this class of models focused on model $\{\phi_t, p_t\}$ and emphasized estimators that existed in "closed form." That is, the calculus was used to take first partial derivatives of the log-likelihood function, set these to zero, and solve the resulting set of equations to get computable estimators "formulae." The parameter being estimated was on the RHS and the data could be substituted for the notation on the LHS and a numerical value produced.

There are problems with this approach:

1. Lots of notation is needed, and this is a bother to students trying to cope with the mass of symbols, subscripts, Greek characters, etc.

2. The notation has changed over the years as close relationships were found with the band recovery models.

3. Many (most) models of real biological data do not have closed form estimators (they do not exist).

4. In some ways, the focus on the closed form estimators hides the important concept of a log-likelihood and its parameterization, the notion of parameter values that maximize this function, etc.

5. While some estimators are "computable" and exist in "closed form" they are often tedious and error-prone if done by hand (e.g., Jolly 1965). All such calculations are done by computer anyway, thus we are teaching FW663 with relatively little emphasis on closed form estimators and all the associated notation required. Our approach puts a premium on the numerical maximization of the log-likelihood and the shape of this function.

6. For use in estimating population size at each time j (i.e., N_j), this model is subject to substantial bias in many cases. Bias arises from heterogeneity of individuals and behavioral response to trapping and handling. In addition, (small sample) bias may arise from small sample size in either R_i or m_{ij}. Curiously, $\hat{\phi}$ trends to be relatively free of bias. Conceptually, $N_j = \hat{N}_j \cdot b$ (where b is bias). But ϕ_j is defined as N_j/N_{j+1}. So, the estimator $\hat{\phi} = N_j \cdot b/N_{j+1} \cdot b$ and the bias term in the numerator and denominator tend to cancel out. Thus, the estimator $\hat{\phi}$ generally has substantially less bias that the estimator $\hat{N}$.

Some questions to ponder:

1. What if there is some individual variation among individuals ("individual heterogeneity")? What effect might this have on $\hat{\phi}$ or $\hat{p}$?

2. What if the time period between captures were unequal? What would this do to $\hat{\phi}$? That about $\bar{\phi}$?

3. What would you expect about precision if the R_i were large (say 1,100) but the p_i were very small (say 0.02)?

4. What treatments might be done with experimentation with marked populations.

5. People interested in estimation of population size might want to study the "robust design" variation of the Jolly-Seber model.

6. Given a data set, what model should be used for making inferences from the sample data to the population of interest? How is one to decide upon a model; *MARK* allows the researcher to build many models with ease – which one(s) should be used?

Key References

George Jolly's (1965) paper is still an insightful paper. Lebreton et al. (1992) and Burnham et al. (1987) and Pollock et al. (1990) provide background reading. Also, see George Seber's 1982 book. Cormack's (1964) papers and Seber's (1965) paper are more difficult to read. See White et al. (1982) for photos and background of these people.